Lexicographical Explorations of Neologisms in the Digital Age. Tracking New Words Online and Comparing Wiktionary Entries with 'Traditional' Dictionary Representations

Creese, Sharon

Award date:
2017

Awarding institution:
Coventry University

Lexicographical Explorations of Neologisms in the Digital Age. Tracking New Words Online and Comparing *Wiktionary* Entries with 'Traditional' Dictionary Representations

By

Sharon Creese

January 2017

A thesis submitted in partial fulfilment of the University's requirements for the Degree of Doctor of Philosophy

Low Risk Research Ethics Approval

Where NO human participants are involved and/or when using secondary data - Undergraduate or Postgraduate or Member of staff evaluating service level quality

Project Title

An Exploration into the Relationship between Lexicography and Language Growth in the Age of the Collaborative 'Wiki' Dictionary

Principal Investigator Certification

I believe that this project **does not require research ethics approval**.	X
I confirm that I have answered all relevant questions in the checklist honestly.	X
I confirm that I will carry out the project in the ways described in the checklist. I will immediately suspend research and request a new ethical approval if the project subsequently changes the information I have given in the checklist.	X

Principal Investigator

Name: Sharon Creese.....

Date: 11/10/2013.....

Student's Supervisor (if applicable)

I have read the checklist and confirm that it covers all the ethical issues raised by this project fully and frankly. I confirm that I have discussed this project with the student and agree that it does not require research ethics approval. I will continue to review ethical issues in the course of supervision.

Name: Hilary Nesi.....

Date: 24/10/2013.....

Acknowledgements

There are a number of people to whom I am deeply indebted for their help and support during the course of this research project

Firstly my supervisor, Professor Hilary Nesi, who has been a constant source of support, insight and guidance on this long journey. Despite my endless questions and unwavering desire to control the uncontrollable, she has remained at my side, guiding me, nurturing my instincts, and helping me to become a better researcher, writer and person. I cannot thank her enough for her time, patience and support.

Professor Sheena Gardner and all of the staff, lecturers and colleagues who have taught me so much during my time at Coventry University.

Dr David Nixon, without whose trust and support this project would not have been possible. And finally Julia and David Hardie, whose unwavering support has helped me to find the strength to keep on working.

Thank you all.

I dedicate this work to Freda Creese, who reminded me
(in her own inimitable way!) that anything is possible.

Abstract

This thesis explores neologisms in two distinct but related contexts: dictionaries and newspapers. Both present neologisms to the world, the former through information and elucidation of meaning, the latter through exemplification of real-world use and behaviour.

The thesis first explores the representation of new words in a range of different dictionary types and formats, comparing entries from collaborative dictionary *Wiktionary* with those in expert-produced dictionaries, both those categorised here as 'corpus-based' and those termed 'corpus-informed'. The former represent the most current of the expert-produced dictionary models, drawing on corpora for almost all of the data they include in an entry, while the latter draw on a mixture of old-style citations and Reading Programmes for much of their data, although this is supplemented with corpus information in some areas.

The purpose of this part of the study was to compare degrees of comprehensiveness between the expert and collaborative dictionaries as demonstrated by the level and quality of detail included in new-word entries and in the dictionaries' responsiveness to new words. This is done by comparing the number and quality of components that appear in a dictionary entry, both the standardised elements found in all of the dictionary types, such as the 'headword' at the top of the entry, to the non-standardised elements such as Discussion Forums found almost exclusively in *Wiktionary*.

Wiktionary is found to provide more detailed entries on new words than the expert dictionaries, and to be generally more flexible, responding more quickly and effectively to neologisms. This is due in no small part to the way in which every time an entry or discussion is saved, the entire site updates, something which occurs for expert-produced online dictionaries once a quarter at best.

The thesis further explores the way in which the same neologisms are used in four UK national newspapers across the course of their neologic life-cycle. In order to do this, a new methodology is devised for the collection of web-based data for context-rich,

genre-specific corpus studies. This produced highly detailed, contextualised data that not only showed how certain newspapers are more likely to use less-well established neologisms (the *Independent*), while others have an overall stronger record of neologism usage across the 14 years of the study (*The Guardian*).

As well as generating findings on the use and behaviour of neologisms in these newspapers, the manual methodology devised here is compared with a similar automated system, to assess which approach is more appropriate for use in this kind of context-rich database/corpus. The ability to accurately date each article in the study, using information which only the manual methods could accurately access, coupled with the more targeted approach it can offer by excluding unwanted texts from the outset made it the more appropriate approach.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BNC	British National Corpus
DDEB	Dictionary Date of Entry Batch
FAQ	Frequently Asked Question
GAS	Google Advanced Search
IPA	International Phonetic Alphabet
KE	Knowledge ecosytem
MW	Merriam-Webster Dictionary
NRS	National Readership Survey
ODE	Oxford Dictionary of English
ODO	Oxford Dictionaries Online
OEC	Oxford English Corpus
OED	Oxford English Dictionary
OUP	Oxford University Press
POS	Part of Speech
RTBF	Right to be Forgotten
SAMPA	Speech Assessment Methods Phonetic Alphabet
SEO	Search Engine Optimisation
SkE	Sketch Engine
SRP	Search Results Page
URL	Universal Resource Locators

WAC	Web-as-corpus
WBC	WebBootCaT
WFC	Web-for-corpus

Chapter 1 Introduction

1.1 Introduction

The computerisation of lexicography dates back to the 1960s, when a database was designed to 'categorise and sort units of dictionary information' (Nesi 2009: 458). Since that time, computers and dictionaries have become more and more intertwined, with the development of computerised corpora and electronic dictionaries (from hand-held devices and CD-ROMs, to dictionary software and most recently online dictionaries (Ibid: 460-2, 467, 472)).

For me, however, the true 'digital age' began with what is widely termed 'Web 2.0'; the move to interactive web technologies, allowing for more 'communicative interactivity, flexibility, social connectivity, user-generated content, and textual creativity' (Danesi 2016: 67, 282). All of this enables much higher levels of participation and collaboration in all aspects of online information sharing.

The idea for the 'World Wide Web' dates back to 1989, when it was first floated by Tim Berners-Lee of CERN (the European Organisation for Nuclear Research). However it was only in 1993 when CERN ceded its rights to royalties on web documents, that the Internet became an option for storage of electronic dictionary data (Nesi 2009: 472). The term 'Web 2.0' was coined in 2001 (Neuman, Nave and Dolev 2010: 58), and indicated a new age of interactivity online. This was around the same time that *Wikipedia* and *Wiktionary* were launched (2001 (Bryant, Forte and Bruckman 2005: 1) and 2002 (Meyer and Gurevych 2012: 261) respectively). It was also when the *Oxford English Dictionary* (*OED*) first appeared online, and when editors decided that new words entering the dictionary should be published in the updates to the online version, rather than being published all together in a separate volume, as had previously been the case (Weiner 2009: 400-2)). Over the next few years, social media sites like Facebook and Twitter appeared, and quickly gained users. In around the same period, UK national newspapers launched online versions), containing the same material as the printed editions, but with a more dialogic approach, enabling and inviting interaction with their readers through blogs, reader comments, and an

increasing number of social media platforms (Facchinetti 2012: 147, 159-60). *The Guardian,* for example, is well-known for its relationship with its readership.

This change led to many new words and phrases entering the lexicon. In the context of this study, 'lexicon' is defined as the total language set available to speakers of English; this includes all of the words included in all dictionaries of English (general and specialised) plus any new words entering the language either through word formation processes (see 1.2.2.1) or through borrowings from other languages. Minkova and Stockwell claim that 'at least 1,000 new and revised entries' feature in each quarterly update of the *Oxford English Dictionary* (2009: 5). Words such as 'google' and 'tweet' appeared, while others gained new meaning. 'Friend' and 'inbox' became verbs, while 'Twitter' and 'like' were nominalised to become nouns (although 'like' retained and even expanded its verb status, with the addition of the new meaning of pressing an electronic (usually 'thumbs up') button to signal agreement in a range of social networking applications).

Mindful of these technological changes, I set out to conduct a lexicographical study, exploring new words in the 'digital age'. I aimed to examine entries for a set of 34 neologisms in four top expert-produced English dictionaries, and to compare them with corresponding entries in collaborative dictionary *Wiktionary* – in my view the most comprehensive of the collaborative dictionaries, and one of the products of this new 'digital age'. I also sought to draw an accompanying picture of these same neologisms in 'real-word' usage, and I chose to do this by tracking their behaviour and use in online versions of several UK national newspapers between 2000 and 2014.

Newspapers were chosen as the best source of data on real-world neologism use because they are produced daily, meaning they can better keep pace with language change than other written materials, and because they are aimed at a broad cross-section of the population. A wide variety of income levels, educational stages, and social groupings can be reached by newspapers. There were additional reasons for choosing newspapers as the medium for showing these neologisms in the 'real world' however. There exists between new-words-in-dictionaries and new-words-in-newspapers a web of connections comprising a complex inter-relationship. When we

examine these interconnections we can see that new words move both from newspapers towards dictionaries, and from dictionaries towards newspapers. While I will not be exploring these relationships here, they are why I believe that the two halves of this study make a useful and cohesive whole.

- Many new words are actually created by journalists, originally for use in newspapers (Renouf 2007: 70), but later spreading into wider use and, it was my belief, into collaborative and later expert-produced dictionaries
- New words and word formations created elsewhere often first come to language users' attention in newspapers (Fischer 1998: 68-9)
- Much of the information in corpora used to create dictionaries (such as the *British National Corpus*) in fact comes from newspapers or other media outlets (Grefenstette 2002: 201)
- It seems likely that newspapers will be one of the publication types referred to in the attestation process which ultimately leads to a new word being accepted into a dictionary. A decade ago, the focus for both *Oxford English Dictionary* and *Merriam-Webster* was printed material (Mitchell 2008: 33) however today this also includes online information, which as mentioned above, can also mean newspapers[1].

The issue of what comprises a neologism is a thorny one however. As Kerremans points out, many lexicographers and linguists writing on the topic do not provide a precise definition of the term, instead considering it to be self-explanatory (2015: 29). Kerremans herself never actually defines what she means by a 'neologism', instead adopting dictionary definitions such as 'new words' or 'new senses or uses of existing words' (Ibid: 27). Kerremans' methodology, however suggests that she views neologisms as being new words which have yet to enter a dictionary (see 2.4). This viewpoint is shared by Fischer, who actually states that a word is considered new if it has not appeared in a dictionary (1998: 3).

[1] See for example https://en.oxforddictionaries.com/explore/oxford-english-corpus

The ‘neologisms’ discussed in this study are words which have been deemed ‘new’ by the *NeoCrawler* (EnerG, n.d.) neologism identification, tracking and analysis software, and which also:

- do not yet appear in any of the five dictionaries used in this study, or
- have entered one or more of the selected expert-produced dictionaries and/or entered collaborative dictionary *Wiktionary,* since 2000 and
- do not as yet experience consistent year-to-year usage in the four UK national newspapers included in this study.

The objectives of this study were as follows:

1. To compare degrees of comprehensiveness (defined as ‘being of large content or scope’[2], in this case relating to the number and quality of dictionary components as established by standard lexicographical practice) in the entries provided for new words in expert-produced dictionaries with those in collaborative dictionary *Wiktionary*
2. To track neologism appearances in UK news media in order to compare usage and behaviour in different newspapers at different stages in the neologic life-cycle. ‘Neologic life-cycle’ is a term coined specifically for this study and refers to the period during which words are said to fit the definition of ‘neologism’ above. However it does not refer to the life-cycle of an individual word, but the generic life-cycle of neologisms. Hence it is examined through tracking of words which have not yet entered a dictionary, words which have recently entered, and words which have been present for several years (see 3.4.4). By examining all three categories, a general picture can be established showing how we might expect individual neologisms to behave over the same timeframe.

[2] See for example http://www.oed.com/view/Entry/37859? and http://www.oed.com/view/Entry/37861?redirectedFrom=comprehensiveness&

In examining the differences between the entries in the expert-produced dictionaries and in *Wiktionary* this study also considers how these are distributed across three different dictionary creation formats (see 3.4.1):

- 'corpus-based' (which I define as dictionaries promoted as being created using mainly corpus data, for example *Oxford Dictionary of English*)
- 'corpus-informed' (defined as dictionaries promoted as being created using mainly old-style Reading Programmes and citations, for example *Oxford English Dictionary*)
- 'collaborative' (defined as dictionaries promoted as being created through collaboration with and between users, for example *Wiktionary*).

This is an important question. If results show that either of the second two methods of dictionary creation is more comprehensive than the 'corpus-based' model, this might begin to prompt questions about the future of dictionary-making.

In order that all of the necessary contextual information be gathered about each appearance of each neologism in all of the newspapers chosen for this study, a new methodology for corpus data collection was devised, since existing methods (automated, to gather maximum data) are not designed to collect the required level of contextual information. In order to explore the suitability of this new methodology for conducting future context-rich 'genre'-specific language studies like this one, this new methodology was compared with the most recently written-up automated program aimed at identifying and monitoring new words: the *NeoCrawler* (see 2.4). In relation to such future studies, 'genre' is defined here as referring to 'different communicative events which are associated with particular settings and which have recognised structures and communicative functions' (Flowerdew 2013: 138). Many of the different sections and articles of newspapers clearly meet this definition, for example sports or financial writing. However I am conscious that newspapers as a whole do not, since they act as a medium to bring together all of these different writing styles. Thus where the word 'genre' is used in relation to newspapers it is done so only in the absence of a better, all-inclusive term.

All of this gave rise to a third objective of this study:

3. To consider whether neologism use and behaviour in the media can be best explored through the use of new manual or existing automated corpus data collection techniques.

It should be made clear, however, that while this new methodology is put forward as a 'corpus data collection' tool, and it is planned that in the future this will be its main purpose, the tracking of neologisms in newspapers in this study was not intended to produce a corpus, and this is not a corpus linguistics study. A corpus enables researchers to make a range of different linguistic queries (Hunston 2002: 3), but in this lexicographical study the database was designed for the purposes of examining a particular set of neologisms within a particular timeframe.

As this new methodology offers future researchers a tool for producing more contextually nuanced genre-specific corpora (or databases) than is possible with programs such as the *NeoCrawler*, it is therefore considered one of the key contributions of this project to academic study. It should be pointed out, however, that although this study tracks neologism use in newspapers as well as examining representations of new words in dictionaries, it does not undertake the task of finding new words which have previously gone unnoticed.

1.2 Background

Before presenting a detailed review of the literature currently available on the topics covered in this study (see Chapter 2), let me first begin with a brief discussion of factors which I believe will be relevant to the reader.

1.2.1 Motivation and Initial Ideas for this Study

The current study developed from an initial idea to explore the relationship between lexicography and language growth. One way to do this would be to examine the development and spread of one of the key agents of such language change – neologisms – in both UK national newspapers and in dictionaries. Having previously worked as a journalist and as an editor for a dictionary publisher, I felt I was in a particularly strong position to take on such a piece of research, since I had retained a clear understanding of both the requirements of professional journalistic writing and newspapers' customary approaches to new and unusual words, as well as a keen interest in how the dictionary marketplace works, how it might develop over the coming years, and in particular how non-traditional dictionary formats might fit into that revised landscape. I was especially interested in the role, growth and potential future of one such 'non-traditional' dictionary format, that of 'collaborative' dictionaries, since they were clearly growing in popularity, and *Wiktionary* in particular had reached the point where it looked and functioned very similarly to expert-produced dictionaries. I was interested to see how this affected its relationship with new words, whether it was, in fact, 'as good as it looked' and in addition, what its presence in the marketplace might mean for 'traditional' dictionary publishers both now and in the future. As mentioned in 1.1, a central element of modern-day lexicography is the use of a corpus, 'a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research' (Sinclair 2004: 23). Yet *Wiktionary* and other collaborative dictionaries do not use corpora and as I explored the possibilities of a research project such as that outlined above, I further began to wonder how the absence of a corpus affected *Wiktionary*, and whether it might even be the case that a corpus could become more of an impediment than an advantage in the modern digital age.

As I began to investigate these ideas, it further became clear that *Wiktionary*'s provision of a date of inclusion for <u>every</u> new word on its site, would be crucial to a study such as this. On discovering that words tend to enter *Wiktionary* before expert-produced publications, due to the former's more relaxed inclusion criteria (see 3.4.2),

I decided that the *Wiktionary* entry date would therefore be taken as indication of when a word had entered what we generically refer to as 'the dictionary'. (However, if a word entered the *Oxford English Dictionary* (the only other dictionary carrying any date information at all) before *Wiktionary* then that would be the assigned date.)

Having established *Wiktionary* as the indicator of the date of entry of a new word into 'the dictionary', it was clear that the ability to collect similar information about a neologism's appearance in a newspaper would be just as crucial. Indeed, it would not only be publication date, but a range of additional 'contextual' features that would be required from newspapers in order to paint the kind of in-depth picture of neologism usage that I sought. This would require a significant change in methodology from the norm, and so, as my thinking around these core issues became clearer, the thrust of the study became more targeted, developing into a lexicographical exploration of new words in dictionaries and in newspapers, featuring the development of a new manual methodology for the collection of web-based data.

1.2.2 Theoretical Backdrop to this Study

It is customary in a study such as this, to lay out the theoretical backdrop against which the current research is conducted. This allows the author to position his/her project within the bounds of contemporary theoretical thinking, and provides the reader with a theoretical context within which to read and evaluate the research. In the case of the current study, there are two fields in which one would hope to be able to provide this kind of theoretical backdrop: lexicography and neology. In reality, however, there is no lexicographical theory which applies to the current study, and hence there is no context which can be provided to the reader outside of the review of literature provided in Chapter 2.

Similar problems arise when we turn to neology, or the study of new words. While studies exist on various aspects of neologisms, no theory appears to have been advanced on their creation, development or use. Indeed the difficulties of terminology which will be discussed in 2.2.1 make the development of such a theory problematic,

since there does not appear to be a consensus on the meaning of the key concepts behind the terms ‘productivity, ‘creativity’ or ‘life-cycle’.

As a consequence, for neology as for lexicography, I must rely on recent literature to set the scene for the current study: see Chapter 2, *Literature Review*.

1.2.2.1 Morphology and Neologisms

While it has not been possible to establish a theory for the study of neologisms, it is important to present background information on the topic of morphology, in order to better understand the word formation processes (WFPs) behind the neologisms in this study. This will prove useful in the discussion of the selection of neologisms from the *NeoCrawler* list of potential candidates (See 4.2 and its subsections). In order to understand the various word formation processes, I begin with an explanation of the ‘smallest unit that has meaning or serves a grammatical function in a language’: the ‘morpheme’ (Katamba 1994: 32). To be a word, a lexical unit must contain one morpheme (thus being ‘monomorphemic’, for example ‘stack’) or more than one morpheme (thus being ‘polymorphemic’, for example ‘helpful’, made up of ‘help’ + ‘-ful’) (Carstairs-McCarthy 2002: 17).

The term ‘helpful’ is useful in illustrating two of the three different types of morpheme present in English: the ‘root’ and the ‘affix’. The root is the base form of a word, structurally usually a ‘free’ morpheme, meaning that it can stand alone (here ‘help’), while the affix is the ‘bound’ morpheme, which only becomes a word when joined with a root form (here ‘-ful’) (Katamba 1994: 40-1, 54-5; Carstairs-McCarthy 2002: 20-1). A third morpheme is the ‘combining form’, words containing more than one root morpheme, such as compounds (words which are made up of two or more existing words joined together, at least one of which is a root morpheme) (Ibid: 21). Carstairs-McCarthy includes ‘blends’ (comprising parts of words joined together, rather than complete ones), ‘phrasal words’ (‘items that have the internal structure of phrases but function syntactically as words’) and ‘acronyms’ (words featuring just one letter from each of the constituent parts) within the category ‘compounds’ (Ibid: 59, 65, 67-8). However, I, like Minkova and Stockwell (2009: 14-16) consider ‘blends’ and

'acronyms' to be additional word formation processes (along with others such as 'clipping' (cutting off part of a word and retaining the rest, such as 'fav' from 'favourite'), 'back-formation' (using the clipped off element as a word in its own right, since the part which has been removed is recognised as an affix, for example 'edit' from 'editor') and initialisms ('acronyms' in which the individual letters are pronounced, such as *OED* for *Oxford English Dictionary*).

All of these processes fall into the category of 'derivation', where new words are created out of existing words or morphemes. The alternative to this in terms of morphemic processes is 'inflection', which involves the addition of an affix to ensure that the word fits the required grammatical context, for example the addition of an '-s-' to produce the third person singular form of a present tense verb, such as 'see+s' (Katamba 1994: 58-9).

There is one further derivational method which, although referred to as 'zero derivation', is actually another example of the derivation of new words from existing ones. This process is also known as 'conversion', the method by which words change to, or develop a new meaning in a different word class (part of speech). Very often these changes are to/from nouns and verbs (for example the word 'jump') (Katamba 1994: 70-1) however they can also involve other parts of speech, such as nouns born of adjectives, or adverbs born of adjectives (Carstairs-McCarthy 2002: 45-50). In the case of conversion there is no visible change to the word (such as 'hope' or 'fear') hence the 'zero derivation' label (Ibid: 48).

One word formation process falling outside the 'inflection' / 'derivation' dichotomy is what Minkova and Stockwell refer to as 'creation *de novo*', which simply means that the word has been created from scratch (although this is claimed to be rare) (2009: 12-13). One example is the noun 'google'; beginning life as a trade name. It is possible that 'Google' is related to 'googol' meaning ten to the power of a hundred and thus representing the vast amount of information available via the Internet (Ibid). Many *de novo* words start as trade names and gradually develop into generic terms, such as 'xerox' for 'photocopy' and 'hoover' for 'vacuum'. Another route for new words to enter the language is as 'borrowings' or 'loan words' from other languages. Among

these are ‘catachrestic’ loanwords, or those which have been introduced into English in order to ‘fill a lexical gap opened up by the introduction of a novel object, concept or idea’ from the source language (Barrs 2015: 372). Thus the introduction of the idea creates the need for the word to name it.

For analytical purposes, in a study such as this where we are examining words which have already been formed, it is important to understand how to reverse the process and ‘deconstruct’ these words, breaking them down in their constituent parts. Since morphemes are the smallest units into which words can subdivide, only words with two or more morphemes can be divided, or ‘segmented’ in this way. Ginzburg et al (1979: 90) identify three levels of ‘morphemic segmentability’:

- complete
- conditional
- defective.

In the many words featuring ‘complete morphemic segmentability’ the morphemic structure is clear enough to make the individual morphemes easily identifiable (Ibid). ‘Conditional morphemic segmentability’ describes those words where breaking down into individual morphemes is difficult due to semantic issues. What this means in practice is that one or more of the component parts is not, in fact, a true morpheme; instead these are termed ‘pseudo-’- or ‘quasi-morphemes’ (Ibid).

‘Defective morphemic segmentability’, meanwhile ‘is the property of words whose component morphemes seldom or never recur in other words’ (Ibid). One of the morphemes making up the word is considered ‘unique’, indicating that it derives its meaning directly from the other morphemes around it; without those (that is, in a different word or linguistic context) it would have no meaning at all (Ibid: 90-91). One example of a word demonstrating ‘defective morphemic segmentability’ is ‘cranberry’: without the morpheme ‘-berry’, ‘cran-’ has no meaning. The word therefore cannot be successfully broken down (Ibid: 91).

1.2.3. British National Newspapers

This section provides a snapshot of the British national newspaper marketplace, out of which the four newspapers in this study were chosen. It also offers a brief summary of the key points in the linguistic history of such newspapers.

The 'British national press' is defined by Cole and Harcup (2009: 19) as 'those newspapers published in London and readily available across the UK'. There is a dominance to the national press in this country which is not found in many other European nations or in the USA. This is in part due to those countries' larger size, making overnight delivery of newspapers from a central location to outlying locales impossible. There, regional newspapers rise up to fill the void, each centred around a particular major city (Ibid: 19-20), for example the *New York Times* or the *Washington Post*. The dominance of national newspapers has been the case in the UK for well over a century, with rail links originally responsible for transporting the daily newspapers, followed by a move to roads in the 1980s (Ibid) and the current move to even faster methods of delivery offered by the Internet. Of course things did not begin this way. England's first newspaper was *The Oxford Gazette* (later renamed *The London Gazette*), which began publication in the mid-1600s (Fries 2012: 54; Brownlees 2012: 2-3). This followed the newsheets and pamphlets which had previously carried the news, and prior to that the coffee houses which had reported such information in person (Ibid; Cole and Harcup 2009: 61). The first daily paper was *The Daily Courant*, which began in 1702 (Ibid).

Today, the British national newspaper marketplace is broken into three main categories: 'quality' 'broadsheet' newspapers comprising the *Telegraph*, *Times, Guardian, Independent* and *Financial Times,* 'mid-market' (tabloid size) newspapers – the *Mail* and the *Express*, and the 'red-top' tabloids comprising *The Sun, Mirror* and *Star* (Ibid: 20) (although all of the tabloids are often joined together in the same category). Similar stratification is seen in the Sunday newspapers (Ibid), although there will naturally have been some changes since the closure of the *News of the World* in 2011 as a result of the phone hacking scandal. In the current study, data from each newspaper includes both weekday and Saturday/Sunday editions. Thus

information on neologisms listed as appearing in the *Mail* includes that on neologisms in the *Daily Mail* and the *Mail on Sunday*. This is not always the case in the literature on newspapers, where it is not always clear whether an author is writing about both weekday and weekend editions. Those providing readership figures appear to prefer to analyse separately, and this seems to be because different patterns of readership are found in the two categories. Duffy and Rowden, for example, of Social Research Institute Mori, include only weekday publications in their 2005 review of newspaper readership figures. Since I am not looking at newspapers in terms of readership figures, but in terms of neologism frequencies, it is not necessary to observe this distinction. Thus I incorporate both weekday and weekend usage in order to show how many times new words appear in the newspaper as a whole.

The *Telegraph* was the first newspaper to go online, in 1994 (Facchinetti 2012: 147), meanwhile the *Independent* (including the *Independent on Sunday*) was the first to commit exclusively to electronic publishing, ceasing publication of its print version in March 2016. For nine weeks in early 2016 an additional national newspaper was published, the *New Day* (Trinity Mirror 2016). Claiming to be 'politically neutral' and aimed at a 'time-poor' audience, the *New Day* surprisingly had no website, although it did have 'a social media presence'[3]. Reviewing the newspaper myself, I felt it appeared to be trying to adopt many of the interactive characteristics of social media and apply them to a print format; this seemed unlikely to succeed. For example pages were reserved for reader feedback (from social media platforms such as Twitter) to be printed out. It also struck me as highly odd to be so pro-digital-media, and yet to intentionally fail to provide one of the key vehicles for digital media, a website.

The *New Day's* approach to social media demonstrated a key aspect of newspapers, that they are a product of the culture in which they are published. Indeed Reah states that 'newspapers are cultural artefacts', and that the language they use also reflects the culture which gives rise to them' (1998: 54).

Variation in the language of different types of newspaper began to really take hold in what is now known as the era of 'New Journalism', in the early 20th Century. 'New

[3] http://www.bbc.co.uk/news/uk-36209318

Journalism' involved major changes in the 'content, layout and style' of newspapers, with language moving away from an educationalist tone to one more representative of its readers (Bös 2012: 98). Along with 'New Journalism', 'tabloidisation' saw the beginning of a clear distinction between tabloid newspapers, with shorter stories and lots of captions, aimed at the working classes, and non-tabloids offering hard news aimed at the higher echelons of society (Bös 2012: 101-105). Different elements in the newspaper featured different linguistic styles, for example 'news interviews' were used to introduce language into the newspaper which was more familiar to readers' own speech communities (Ibid: 100).

These changes in the newspaper marketplace gradually developed into the 'tabloids versus broadsheets' structure we see today. Journalism is now considered to be information, and 'the language of news is supposed to be first and foremost factual' (Facchinetti 2012: 145). While information is available on the language of newspapers over time, thanks in part to a number of corpora such as the Rostock Historical English Newspaper Corpus (from 1700 to the present day), the ZEN Zurich English Newspaper Corpus (late 17th and 18th Century newspapers) (Fries 2012: 51) or the Reuters Corpus (Facchinetti 2012: 171), none of the references to these corpora appear to mention neologisms. For that we must look to the studies discussed in the Literature Review.

Of course the most recent change to the world of journalism is the addition of social media, as noted with reference to the *New Day* above. Most newspapers now also produce a blog, and these 'have impacted significantly on the conduct of journalism' (Danesi 2016: 272). Of the four newspapers used in this study, only one integrates its blog offering with the electronic version of the main newspaper: *The Guardian*. The rest provide their blog at a separate web address. It will be interesting to see over the coming months and years whether other newspapers follow *The Guardian*'s lead in this, resulting in a complete integration of their social and traditional media offerings. If they do, this might have some effect on the linguistic style of the newspaper, and the use and development of neologisms within it.

1.3 Thesis Outline

As mentioned in 1.1 the purpose of this lexicographical study is to explore dictionary representations of new words in the 'digital age', comparing entries in expert-produced dictionaries with those in collaborative dictionary *Wiktionary*. Accompanying this is a picture (drawn through tracking of neologisms in the media) of how these new words behave and are used in the real world, specifically in newspapers, which very often give rise to the corpora on which modern-day expert-produced dictionaries are based.

Chapter 2 *Literature Review* comprises a detailed review of the literature currently available on key topics for this study, as well as identifying significant gaps in the literature which the current study seeks to fill. The lack of writing on the relationship between neologisms and lexicography is revealed, along with the lack of any studies comparing entry information in 'corpus-based-, 'corpus-informed' and 'collaborative' dictionaries. Finally this chapter discusses the lack of any significant body of literature on methodologies for working with neologisms. Chapter 2 further reviews the literature concerning lexicography, social media and neologisms and lexical creativity as they appear in dictionaries and in newspapers. The use of corpora as a tool in dictionary-making is also discussed, and the *NeoCrawler* neologism identification, monitoring and analysis program is introduced.

Chapter 3 *Methodology Part 1: Laying the Groundwork* presents key elements of the project, including dictionaries used for comparing representations of neologisms, and newspapers used to track usage of those new words. It outlines the processes leading to the development of the new methodology (execution of which will be covered in Chapter 4) and discusses the *NeoCrawler*, whose automated processes are compared to the new methodology created here. This chapter further lays out the methodological framework for the study, and discusses the importance of research validity and reliability, and how they can be achieved.

Chapter 4 *Methodology Part 2: Data Collection and Analysis* discusses how the elements of the project introduced in Chapter 3 were executed. Methods include

comparison of expert-produced versus collaborative dictionary entries for neologisms, media tracking to demonstrate usage and behaviour of these new words in the real world, and comparison of automated versus manual methods of corpus data collection.

The process of identifying neologisms for the study is explained, along with the different methodological components which were tried and tested before the final process was put in place. Collection of newspaper data using new manual methods such as 'pre-screening' of search engine results and 'advance exploration' of target web pages is described, along with the collection of dictionary data through the breaking down of entries from different types and formats of dictionary into their component parts. Methods of analysing these two strands of data are then explained.

Chapter 5 *Findings and Discussion* presents the results of these analyses, and discusses them in light of the Research Questions presented in section 3.9 of Chapter 3. Throughout this chapter, as throughout the entire thesis, the objectives laid out in 1.1 remain central to the discussion:

1. To compare degrees of comprehensiveness in the entries provided for new words in expert-produced dictionaries with those in collaborative dictionary *Wiktionary*
2. To track neologism appearances in UK news media in order to compare usage and behaviour in different newspapers at different stages in the neologic life-cycle
3. To consider whether neologism use and behaviour in the media can be best explored through the use of new manual or existing automated corpus data collection techniques.

Areas of discussion arising from the findings of this research project include the contrasting representations of neologisms found in expert-produced versus collaborative dictionaries, and in dictionaries which are 'corpus-based', 'corpus-informed' and, again, 'collaborative'. The different components found in different dictionary types are discussed, including the degree to which they match the accepted

standard within the lexicographical field, while additional concepts and components specific only to *Wiktionary*, such as transparency of entries, are also explained. Neologism usage across different newspaper titles is discussed in the second half of Chapter 5, along with discussion of factors influencing the use and development of neologisms, such a socio-economic and cultural changes. The differences in usage based on factors such as article type are also considered.

The exploration of these issues is summarised in Chapter 6 *Conclusion*. This includes the responsiveness of *Wiktionary* to neologisms, including the level of detail included in its entries and issues surrounding methods of data collection for context-rich, genre-specific corpora: manual or automated. This chapter also discusses the implications of the findings of this research project, both in terms of the lexicographical landscape, and of future research.

Chapter 2 Literature Review

2.1 Introduction

The following chapter provides a 'literary context' from which to enter into the 'lexicographical exploration of neologisms in the digital age' which the current study offers. It provides a brief summary and a critical evaluation of the literature which helped to shape the project, and contextual studies which enabled orientation of the current work within the relevant academic landscape. Topics covered in this review include neologisms in the context of dictionaries and newspapers, the changing face of lexicography due to the rise of 'e-' or 'electronic' lexicography, and the role of corpus linguistics as a tool in the making of dictionaries.

Through exploration of this literature, gaps in existing academic research are identified, which the current project seeks to begin to fill. These include a lack of clear methodologies on working with neologisms, comparisons of different dictionary formats ('corpus-based', 'corpus-informed' and 'collaborative') and most importantly, the lack of any comprehensive work on the relationship between lexicography and neology. Two studies (discussed here) do make passing reference to lexicography and neologisms (Renouf 2013, and Kerremans 2015), however they do not provide sufficient detail to be of real use in setting the scene for this element of the research project. Despite this, they are included in this chapter because the former (along with others) is useful in establishing the nature of previous work on neologisms in the media, and the latter is the central work on the *NeoCrawler* neologism identification, monitoring and analysis software program, which generated the neologisms used in my own study. The *NeoCrawler*, which is assessed through an evaluation of the articles written by members of the team which created it (from Ludwig-Maximilians University in Munich) is also the source of the automated methodology against which my own newly created manual methodology for corpus data collection is compared.

The key works critiqued in this chapter are: Moon (2008) and (2009); Fischer (1998); Renouf (2007) and (2013); Meyer and Gurevych (2012); Penta (2011); Abel and Meyer (2013); Kilgarrriff (2013); Grefenstette (2002); Kerremans, Stegmayr and Schmid

(2012); and Kerremans (2015). In the course of this discussion, of course, a number of other sources are also briefly touched upon.

2.2 Neologisms and Lexical Creativity

Neologisms are traditionally created to fill what Janssen (2013) refers to as 'lexical gaps', often occurring in the fields of technology and marketing, with the resulting constructions spreading to the wider lexicon, in many cases through use in electronic and social media. It may be, however, that the rationale behind neologism coinage is changing. When Lehrer talks about the creation of 'clever, trendy, eye-and-ear-catching words', she appears to suggest that these are coined not so much to fill lexical gaps, but simply for the novelty value (2003: 371). Certainly new and witty words appear every day, often in the media, and I would agree that in many cases they are created for linguistic effect. They also serve a communicative purpose, and Francl (2011) argues that this can be aided by the use of innovative and unusual forms. She contends that adopting 'whimsical' new words instead of formal scientific terms offers a greater chance of demystifying science and engaging listeners and readers. She points out that 'more than a quarter of the 45,000 words added to the *OED* in the past decade can be broadly classified as 'science'' (2011: 417-418). The enormity of this figure suggests that there is, indeed, much to demystify.

Many of the 'trendy' neologisms Lehrer refers to could perhaps be termed 'buzzwords', which Neuman, Nave and Dolev differentiate from neologisms through their status as 'fashion words that enter the language and rapidly acquire great popularity'; often these words then fade into obscurity (2010: 58, 67). It seems true that most 'buzzwords' never make it into a dictionary, presumably because they fail to meet the strict inclusion criteria applied by publishers. Mitchell provides a useful description of the process used to determine if a new word is ready for inclusion in *Merriam-Webster*'s dictionaries (see 3.4.2), although the focus on citations in printed

works (as opposed to electronic ones) is now outdated, as *Merriam-Webster*'s own website indicates[4] (Mitchell 2008: 33).

Some of these new words will have been created through the process of 'lexical creativity', a term which appears to possess multiple meanings (to be explored in the following section). 'Lexical creativity' can be loosely understood as the <u>further</u> morphological development of new words and meanings, following their original creation (see for example Renouf (2007); Fischer (1998); Moon (2008)). This usually involves the kind of morphemic word formation processes outlined in 1.2.2.1. Thus new words are created (or borrowed from other languages), and from these, additional new words are derived. These issues of neologisms and 'lexical creativity' are discussed in detail in the coming sections, with regard to articles and chapters by authors such as Renouf (2007 and 2013), Moon (2008), Kerremans (2015) and Fischer (1998).

2.2.1 Issues of Terminology in Articles dealing with Neologisms and Lexical Creativity

As noted in 1.1 whilst defining the term 'neologism', it appears to me (and Kerremans makes a similar point (2015: 29)) that most authors writing on the topic assume that the meaning of the term is self-explanatory. This problem is not confined only to 'neologism'. There are several terms which are used by the authors of the texts reviewed in Chapter 2 which are either given slightly differing meanings, or are not defined at all. This leads to confusion and in some cases contradiction when encountering the same term in a different source.

The following terms are subject to this problem:

- productivity
- creativity
- life-cycle

[4] http://www.merriam-webster.com/help/faq/words_in.htm

Definitions of the terms 'productivity' and 'creativity', and explanations of the relationship between them appears inconsistent across the literature relating to this study. Fischer initially defines productivity as 'the ability of speakers/hearers to produce and understand new words' (1998: 17), which appears to bear little resemblance to Renouf's 2007 definition. There she states:

> Productivity is the term used to refer to the word formation processes wrought upon a lexeme. If a word is 'productive' it means that associated grammatical and derivational variants are being produced (2007: 63).

Neither does it accord with her 2013 definition:

> Productivity is an active, living quality in the language which is realised in the creation of newly-derived and inflected variants of a word across time (Renouf 2013: 189).

However Fischer immediately follows her initial definition with 'in a narrow sense, productivity refers to rule-governed word-formation processes which are carried out by the creation and comprehension of new words' (1998: 17), which appears to more closely fit Renouf's 2007 definition than Fischer's preceding words.

'Creativity' can be confused with 'productivity' since it can indicate the act of applying the word formation processes outlined in 1.2.2.1. This inter-relationship between terms is recognised by Moon when she talks about 'recurrent and productive patterns of usage: something that could be described as systematic [lexical] creativity' (2008: 133). Moon discusses four different types of creativity: 'word meaning, affixation', idiom form and respelling') (2008: 131) and states that 'by "systematic creativity" I mean cases where individual words, phrases and affixes are regularly used in creative ways to produce variations of meaning, including connotation and pragmatic effect' (Ibid: 133). Fischer also mentions the relationship between 'productivity' and 'creativity', but presents them as being so interdependent that it becomes difficult to understand what she is actually trying to say. She firstly introduces 'several types of creative neologisms (i.e. [sic] *shortenings*, *lexical phrases* and *combinations*)' (her

italics) (1998: 1), then goes on to state that '*creative neology* is a term used for word-formation types other than *compounding* and *derivation,* both of which are usually considered to be the only productive word-formation patterns' (again, author's italics) (Ibid: 2). She then adds that in her view 'creative neologisms are also susceptible to productivity, even if not to the same extent as "productive" neologisms (i.e [sic] compounds and derivatives)' (Ibid). Fischer thus places 'creativity' both before and after 'productivity', and gives no clear definition of what the former term means.

She does, however, state that 'creativity is unpredictable and not governed by rules' (Fischer 1998: 17). Renouf more than acknowledges the relationship between 'creativity' and 'productivity'; her 2007 study examines the two processes in concert, within the British broadsheet media. She initially appears to agree that lexical 'creativity' is not subject to rules or regulations, stating that 'creativity is typically thought of as the act or quality of an unpredictable departure from the rules of regular word formation' (70). In journalism she claims that lexical creativity manifests itself in 'punning and other word play, metaphorical extension, willful [sic] error and duplication or usurpation of the role of an existing formation' (Ibid). However she goes on to demonstrate that there are, in fact rules to creativity, referring to 'a clear set of conventions, involving substitution on the basis of phonological, morphological, semantic and other types of similarity as well as allusion' (Ibid: 74). This is somewhat confusing when we see the definition in her 2013 chapter:

> By "creativity" we mean an actual creative rule break; the manipulation of a neologistic word – and particularly a neologistic phrase – to create metaphor, word play or a pun, of the kind which are favoured by journalists (Renouf 2013: 192).

All of this seems to suggest that, while there are conventions guiding linguistic creativity, these operate outside of the standard system of accepted word formation processes. It also indicates that we should consider 'productivity' and 'creativity' to work in tandem. Yet having said this, Moon's 2008 study includes investigation of affixes and blends, Fischer's 1998 book includes acronyms, blends and clippings, and Renouf's 2007 work includes acronyms and blends. This brings us back to the

confusion caused by non-agreement over categories of word formation processes noted in 1.2.2.1. In my view, all of these are standard word formation types. It is therefore my opinion that while lexical creativity may involve more relaxed approaches to the development of new words and/or meanings, it does not necessarily exclude the use of standard word formation practices.

Finally in discussing issues of terminology, Renouf's use of the term 'life-cycle' appears to shift between her 2007 paper on *Tracing Lexical Productivity and Creativity in the British Media*, and her 2013 work *A Finer Definition of Neology in English*. In the earlier work, the 'life-cycle' of words was described as:

> In the most general terms, of birth or re-birth, followed by gentle or steeper upward trajectories in frequency of use and leading to brief or lengthier moments at the zenith of popularity, after which they take faster or slower downward paths, until they reach a stable level of use (2007: 87).

Leaving aside the hedging to accommodate differing results in the data, the key point here is that the 'life-cycle' appears to end when the word ceases to fluctuate in frequency and becomes a stable element of the lexicon. In her 2013 work, however, Renouf appears to have changed her explanation of this 'life-cycle', which now ranges from:

> Its first appearance in our text [newspaper texts from 1989-2011], through its fluctuations in frequency and popularity, to its possible assimilation into mainstream language, and its possible death and re-birth (2013: 177).

The life-cycle of neologisms therefore no longer ends with a stabilisation of frequency fluctuations, but instead carries on to a further, 'final' stage in which the word either 'dies' or is 're-born'/revived. Renouf's apparent change in meaning over her 'life-cycle' may be the result of further development in the author's thinking in response to later studies; it may be that the final stage was simply missed from the earlier article, or it could be that the results of the more recent work required the addition of a new, final stage in the process. The 'trajectory' appears to be the same in both cases; it is the

end point which changes, with a fuller explanation, including specific individual stages, in the 2013 work. This will be discussed in 2.2.3.

2.2.2 Coverage of Methodological Information

In this section, I review and evaluate several works already mentioned in 2.2.1: Renouf (2013) and (2007), Fischer (1998) and Moon (2008). The first of these deals with neologisms in British newspapers while the remainder deal with 'lexical creativity': in newspapers for Renouf (2007) and Fischer (1998), and in dictionaries for Moon (2008).

The most important point to initially highlight is the fact that there is a noticeable lack of methodological information in these sources, and this is one of the contributions that I seek to make with my own research project. How the various studies were conducted is often extremely unclear. Moon makes no reference to methods, aside from referring to her use of the *Bank of English* corpus, and 'three recent British monolingual EFL/ESL [English as a Foreign Language/English as a Second Language] dictionaries' and explaining why these were chosen (Moon 2008: 133-34, 137-38).

Information on methods in Fischer's book is spread across sections and chapters of the work, making them hard to find and making it difficult to draw together any sense of a cohesive methodology. She states early on that she will examine neologisms featuring chosen word formation processes, in samples of 100 drawn from four dictionaries of new words (1998: 23). Each of these is to be examined in light of the concepts of 'motivation', 'institutionalisation' and 'productivity' (Ibid: 20). However the structure of the book means that each word formation process in Part II feels like a separate study, with a mini-methodology, discussion and results section (none of which titled as such) (see for example 'Combinations' (Ibid: 55-63)). Part III of the book (Part I having been a very brief overview of the relevant concepts) seeks to examine these word formation processes within *The Guardian* and *Miami Herald* corpora, however once again information on methods is unclear and inaccessible, being spread throughout a number of different sections (Ibid: 68-182).

Both of Renouf's papers, meanwhile, appear to form part of a single long-term (since 1990 (2007: 62)) project which is 'aimed at investigating aspects of the nature of

neology which are represented in a dynamic corpus' (2013: 179). The project analyses not only neologisms but also new 'word senses', 'sense relations', 'productivity and creativity' (2007: 62-3), and has given rise to numerous 'application' papers/chapters (see for example Baayen and Renouf (1996); Kehoe and Renouf (2002); Renouf (2003)) as well as methodological papers such as Renouf, Kehoe and Banerjee (2005, 2007). It is perhaps because of this that in the two publications reviewed here (2007 and 2013) Renouf gives little more than the most cursory of methodological information. I was left with the sense that the reader is assumed to have read the preceding papers and articles and therefore knows how the research arrived at the current point. Indeed I found it necessary to review some of the preceding sources in order to gain a passing understanding of how the research was conducted (see below).

The methodological information in Renouf's 2007 paper states that data is drawn from UK broadsheet newspapers (comprising *The Times*, *The Guardian*, *The Telegraph*, the *Independent* and the *Observer*), with further data drawn from the Web through use of the team's WebCorp data collection software tool (Renouf 2007: 62). In order to understand this tool, we must consult Renouf and Kehoe (2013: 168), where we learn that it:

> Receives a word or phrase from the linguist and passes this to a commercial search engine such as Google, where it extracts the "hit" pages from the search engine results and processes them to send data back to the linguist in the form of concordances in a choice of formats (Ibid).

In her 2007 paper, Renouf explains that 'a specific time chunk of chronologically sequenced, fresh textual data' is fed 'through a set of software filters which detect novel words as well as new collocational environments of existing words' (63). This is the extent of the methodological information explaining how the study works in the 2007 paper. In the 2013 chapter, Renouf states that she is using the *Guardian* and *Independent* newspapers, and applying 'a combination of linguistic criteria and the lexical-statistical measures created during the AVIATOR … and APRIL… projects' to 'identify the changing status of a neologism in a corpus across time' (Renouf 2013: 179). 'APRIL' is explained in Renouf's 2013 chapter as a tool for parsing (breaking

words down into their component parts) 'each new word at character level, then classifying it grammatically and according to word formation type' (180). The 'AVIATOR', Renouf claims, 'identifies and classifies new words according to simple surface criteria' (Ibid). In an earlier article by Baayen and Renouf on lexical innovation in *The Times* newspaper, the 'AVIATOR' had been said to have been charged with creating an 'automated system for the recording of lexical innovation and change' (Baayen and Renouf 1996: 70). It basically used a series of software filters (designed with lexicographical applications in mind) to locate new words and productive lexical changes such as those discussed in 2.2.1. (A full review of this publication is not provided here as it is a computational linguistics paper focusing on hapax legomena (a word appearing only once within a given context) appearing as a result of morphological productivity (not creative productivity; the words giving rise to these terms are not neologisms) (Ibid: 69-94). As my own research involves tracking use of neologisms in the media rather than finding first instances of their use, the only real relevance of the study therefore lies in the light it sheds on the methodology of later papers focussing on other aspects of the same long-term project).

Two distinctly methodological publications (Renouf, Kehoe and Banerjee (2005 and 2007)) explain how the software used in Renouf's 'application' papers (WebCorp and WebCorp Linguist Search Engine (LSE) was created and built, but this separation into 'methodological' and 'application' in terms of the thrust of publications leads to confusion and problems understanding how results were achieved. This is not helped by the fact that, as far as I can tell, the corpus collected through use of these tools is not available for use outside of the immediate research team.

In summary then, methodological information on working with neologisms is consistently inadequate. It is either missing entirely (Moon 2008), spread throughout a book and hence difficult to bring together and understand (Fischer 1998) or reliant on an apparent assumption that readers are already familiar with the 'methodology so far' (Renouf 2007 and 2013).

2.2.3 Neologisms in the News / Dictionaries

I now review the single study dealing with actual neologisms in the media (as opposed to any associated forms of lexical creativity or productivity (see 2.2.4)): Renouf's 2013 chapter: *A Finer Definition of Neology in English. The Life-Cycle of a Word.* Apart from introducing the reader to the wider neologic project (now midway through its third decade), describing synchronic and diachronic corpora and distinguishing between semantic and grammatical neologisms, this chapter describes the stages Renouf now deems to make up the life-cycle of 'words which are or which have been neologisms in our data' (Ibid: 177-181). It then goes on to present and discuss examples of each one, taken from *The Guardian* and *Independent* newspapers, between 1989 and 2011 (see below), explaining and providing evidence for their categorisation in each case. This classification system to me feels somewhat contrived, presenting a demarcation of stages that I am not entirely confident actually maps onto real-world development of new words (2013: 181). This is perhaps in part because of difficulties surrounding several of the stages due to the issues of meaning mentioned in 2.2.1. The stages proposed by Renouf for a new word's life-cycle are:

- birth/first occurrence in text
- possible increase in frequency of occurrence
- productivity
- creativity
- settling down, assimilation and establishment in the language
- obsolescence, possible death
- possible revival

The choice of neologisms to illustrate this process is, in my view, similarly problematic. Renouf presents case studies of neologisms she claims are passing through each of these stages, yet many of these new words seem to have such a narrow sphere of use

that I cannot help but wonder whether they were chosen simply because there was nothing better available. If so, I am forced to ask just how appropriate the life-cycle stages actually are. For example, I struggle with the idea that 'Eyjafjallajökull' (the Icelandic volcano which downed thousands of aircraft in April 2010, whose name has subsequently been used 'metonymically' to describe that particular eruption (Ibid: 182)) and 'Arab Spring' ('the idea that a new era of political and social enlightenment and liberation is about to begin in the Arab world' (Ibid: 184)) were the only new words 'born' during the 22-year period which were considered worthy of analysis. Few of the neologisms analysed seem likely to be widely used, and more than half of them would seem unlikely to appear in a tabloid newspaper (for example 'graphene', 'FOI' (freedom of information), 'cameron', or 'donkey brown') (Ibid: 182-204). In fact, it seems unlikely to me that some of the new words in the 2013 chapter will stand the test of time. I doubt, for example, that in five years' time 'Eyjafjallajökull' or 'cameron' (in the 'productivity' stage of the life-cycle, producing words related to then Prime Minister David Cameron, such as 'Cameron-ism' (presumably describing something he might say)) will still be in use.

I also question the suggestion that 'video cassette', or particularly 'dialling tone' (currently in the 'obsolescence' phase') were neologisms between 1989 and 2011 (Ibid: 196-7). I recall video recorders (and their cassettes) being well established by the mid-1980s, and indeed *Wikipedia* states that by the end of the decade more than half of British homes owned a VCR (video cassette recorder)[5].

On a more conceptual level, Renouf's chapter fails to acknowledge the impact of other forms of writing – outside of news journalism – on the development and success/failure of neologisms, for example books and social media. This is presumably because it is focussing solely on the data from the newspaper corpus. It also gives no context regarding how a word's first inclusion in the corpus fits with its appearance in the wider lexicon; thus we have no idea just how 'new' the words truly are.

Turning to neologisms in dictionaries, as mentioned in 2.1, in the research literature there is a distinct lack of academic exploration of the relationship between

[5] https://en.wikipedia.org/wiki/Videocassette_recorder

lexicography and neology. Indeed at the simplest level, even references to neologisms in the lexicographical literature tend to be short and relatively superficial. For example in Meyer and Gurevych 2012, which will be discussed in some detail in 2.3.2 and 2.3.4, there is very limited discussion of neologisms within collaborative dictionary *Wiktionary*. What there is shows that high numbers of new words are found in *Wiktionary*, that it 'encode[s] significantly more neologisms than … expert-built lexicons', and that this is possible because the collaborative nature of *Wiktionary* allows for immediate updating of the site (2012: 277). However it should be noted that these 'expert-built lexicons' are not the same as the 'expert-produced dictionaries' used in my own study. Meyer and Gurevych state:

> As expert-built lexicons, we have chosen commonly used computational lexicons, since they allow their data to be automatically accessed in a similar way to *Wiktionary*. This is necessary for a fair comparison between the different types of lexicons' (Ibid: 274).

This means that *Wiktionary* is compared with Princeton's *WordNet* 3.0[6] and *Roget's Thesaurus*[7]. While Meyer and Gurevych claim that this decision was made because these sites access data similarly to *Wiktionary*, I would argue that such a comparison is really not like-with-like. *Wiktionary* is a dictionary (albeit a non-standard one, as will be discussed in 2.3.4) rather than a database or a thesaurus. Information is not organised thematically (per *WordNet* or *Roget's Thesaurus*). Instead, *Wiktionary* operates much as do the electronic versions of expert-produced dictionaries, and hence my decision to compare *Wiktionary* with the latter (for a full explanation of this decision, see 3.4)

References to neologisms are similarly limited in publications about expert-produced dictionaries. Weiner (2009: 401) notes that until the *Oxford English Dictionary* (*OED*) went online in 2000, new words were generally published in separate *Supplements*, since updating the entire dictionary was such a massive task. From 1986, a project entitled *NEWS* (*New English Words Series*) compiled entries for new words and new

[6] http://wordnet.princeton.edu/
[7] http://www.thesaurus.com/Roget-Alpha-Index.html

senses of existing words. When the Second Edition of *OED* (*OED2*) was published in 1989, 5,000 of these were included. Three subsequent volumes of *Additions* were published, in 1993 and 1997, and these were incorporated into *OED2* when it went online (Ibid: 391, 401). Since then, new words have been published online in quarterly updates (Ibid). Former Chief Editor of *OED* John Simpson, who oversaw the inclusion of these new words in *OED2* (Ibid), states that 'neologisms are a window both on language change and continuity' (2007: 147). However in his 2007 article he questions the impact of new words added to a dictionary on the meaning and use of surrounding words (Ibid: 146-8). He argues that while it is right to include new words in dictionaries, it is also important to 'record and analyse changes in the older vocabulary, as these changes tell us more about the new in the same way as neologisms tell us more about the past' (Ibid: 148).

One of the further difficulties with including new words in dictionaries is whether or not they have the 'staying power' to maintain their place. Some may be so new that they qualify as 'buzzwords', and must later be removed when they fade from use (Neuman, Nave and Dolev 2010). The stringent criteria governing acceptance of new words into expert-produced dictionaries are designed to prevent the need for removal (Mitchell 2008: 34). For example Algeo claims that 58% of neologisms included in a lexicographical corpus from the period 1944-1976 did not appear in two key dictionaries examined a decade after the corpus closed (1993: 283). He states that 'just as words are born anew into the vocabulary, so do words die from it (Ibid: 282). Old words gradually fade away, and new words fail to get established, thus aborting or perishing in their infancy. However Algeo does not state whether these words ever entered a dictionary in the first place, only that they appeared in the *Britannica Book of the Year*, in its new words and meanings section (Ibid: 283-4). The usefulness of this study in discussing neologisms in dictionaries is therefore, in my view, limited.

Mitchell reports that Steve Kleinedler of the *American Heritage Dictionary of the English Language* warns there are inherent risks in removing words from a dictionary, since a word which appears to have become obsolete may experience an unexpected return to use (Mitchell 2008: 34). For collaborative dictionaries (and, indeed, electronic forms of traditional works) this is not a problem, since pages can easily be

reinstated, yet Mitchell makes no mention of 'wiki' forms, despite a heavy focus on electronic terminology. The idea similarly does not apply to words entering the *OED*, since once included, they are never removed, enabling the *OED* to 'become part of the historic record of the English language' according to Jesse Sheidlower, *OED* Editor at Large (Ibid: 33-4). This is also the reason why Algeo's study did not include the *OED*, since it is an 'historical dictionary, aiming more than any other at comprehensiveness of inclusion, rather than at a reportage of current use' (Algeo 1993: 283).

2.2.4 Lexical Creativity in the News / Dictionaries

In this section I provide a brief, generalised evaluation of the lexical creativity sources by Moon (2008), Renouf (2007) and Fischer (1998). As the formation of new words is not the main focus of this study, I confine my review of the literature on lexical creativity to studies exploring this issue in relation to lexicography and/or newspapers. As noted in 2.1, I have been unable to find any studies dealing directly with neologisms in a lexicographical context. The closest I have come is a single study on lexical creativity (relating to the creation of new words) and dictionaries (Moon 2008), plus a further study which makes passing reference to lexicography, using dictionaries of new words as the source for a sample of neologisms featuring specific word formation processes (Fischer 1998). While there are other studies on lexical creativity (for example Hanks (2013) *Creatively Exploiting Linguistic Norms*) I have found only two which cover such creativity in relation to newspapers (Renouf 2007 and Fischer 1998).

As discussed in 2.2.1, 'lexical creativity' involves a range of different approaches to the creation and development of new words and meanings. As my own study focuses instead on the behaviour of actual neologisms, making only passing reference to how they were created, this will be a cursory evaluation of these texts, having already discussed the major points of interest. Of the three sources, the only one to investigate lexical creativity and dictionaries is Moon's 2008 work. She examines four kinds of lexical creativity (figurative meaning, word formation, idioms and spelling) in three monolingual dictionaries for learners of English as a second language (2008: 131, 137-8). She argues that these dictionaries might better serve their users if they

included more information on creative aspects of language such as these, particularly since learners may struggle to find accurate information on this anywhere else (2008: 150). She further comments that these learner dictionaries 'perhaps more than larger dictionaries for first-language speakers, are very responsive to language change' (Ibid: 138). Yet she provides no evidence to support this statement. Further, it is my opinion that the majority of learners would find this additional linguistic information confusing and it would detract from their acquisition of basic language skills. This viewpoint is based upon personal observations of students using dictionaries and the belief that were this additional information included, they would struggle to distinguish between what was crucial to their learning and 'fun' word creations (Ibid: 150).

Moon's article offers plenty of examples, mostly taken from the *Bank of English* corpus, to illustrate her points. However a number of these examples seem unsuited for inclusion in a learners' dictionary, since they include terms or concepts which belong to discrete domains that learners are less likely to need to access. For example, the main corpus extract used to illustrate 'Figurative Uses' on page 134 features 'dependency theory' and 'reductionism', both technical terms which I doubt most learners would need to know. From the lack of mention of any electronic versions of the dictionaries, and the comments about lack of space in the dictionary (see for example 2008: 145), it seems that Moon is discussing printed copies of the dictionaries. In this case, I also wonder about the danger of one or more of the creative neologisms falling into desuetude, and since it still appears in the dictionary, making that publication look prematurely 'out of touch' with contemporary language.

As mentioned in 2.2.1, Renouf's 2007 work examines 'productivity' and 'creativity' in British broadsheet newspapers. Newspapers are generally considered to be 'at the forefront of linguistic change' and this makes them 'promising starting points for the study of neology and productivity' (Renouf and Kehoe 2013: 181). Renouf considers a series of words, phrases or 'sub-word morphemes' which have been subject to these processes, and traces their use in *The Times*, *The Telegraph*, *The Guardian*, the *Observer* and the *Independent* between 1989 and 2005 (2007: 62, 64, 70). Renouf states that 'morphologically, once a neologism begins to take hold it typically starts to spawn inflections, derivations and even base forms. Such productivity may occur

almost at once, especially if the word is in the public eye' (Ibid: 66). She then presents a series of case studies of neologisms which have developed in this way, showing the creative neologism and its subsequent derivatives.

While the study is in general comprehensive, the lack of information about differences in usage across the different newspapers in the corpus did leave me with questions. For example I wonder how frequency of these new derivations varies across the different titles and whether any of the productive neologisms appear more in one newspaper than another. In fact the names of the newspapers and even the broadsheet corpus appear just once or twice after the introductory pages (and then generally in captions). Even the section on lexical creativity in journalism (Ibid: 70-71) makes no mention of them. There are a number of concordances each of which is dated but not attributed to a particular publication, and there is plenty of prose explaining the processes at play, but it is still easy to forget that one is reviewing a study of neologisms in the media. Renouf's conclusion states that this chapter has 'sought to make explicit some of the insights about the nature of lexical productivity and creativity in text by employing a research methodology designed to trace lexical activity over time as it was used in UK journalistic text' (Ibid: 86). Sadly I have to say that for me, this objective has not been met. Whilst the outcome of these two factors working in concert has been demonstrated, I still find that individually they remain something of a mystery. This is of course not helped by the confusion caused by the use of these terms in the other sources I review, including Renouf's own later 2013 work. I do not feel that the 'life-cycle' mentioned in the conclusion (Ibid: 87) has been adequately illustrated. I would argue that instead, something akin to a 'neologic tree' formation has been demonstrated, with derivations and new formations sprouting off the central trunk of the original neologism. This I find very interesting, and much clearer than Fischer's (1998) approach to a similar topic, which I now move on to discuss.

Fischer's 1998 work is subtitled 'A corpus-based study of the motivation, institutionalization and productivity of creative neologisms'. As such, one would expect it to be similar to Renouf's 2007 work (above). In fact, it uses one of the same newspapers (*The Guardian*), and covers some of the same time frame (1990-1996). It

also, I believe, uses the same source data (*The Guardian*'s own internal corpus) although as mentioned in 2.2.2, Renouf is never really clear about this. As mentioned in the same section, Fischer's book is broken into three segments; a short introductory section explaining basic concepts (1998: 1-20), Part II comprising 'mini-studies' on creative neologisms, including information on how they are used and develop within several different dictionaries. The neologisms here include examples of acronyms, clippings and blends (Ibid: 21-67). Part III explores use of such examples in *The Guardian* (1990-1996) and also in *The Miami Herald* from 1992 onwards. *The Guardian* was chosen because its corpus was available on CD-ROM (Ibid: 71-2). *The Miami Herald* was probably added for the same reason, since at the time, CD was a new medium for linguistic work (Ibid: 78, 72). Indeed the addition of *The Herald* feels very much like adding an additional source in order to be able to make comparisons. However the comparisons are limited; for example there appears to be no discussion of the differing contexts affecting technological neologisms in the US and the UK, such as 'DAT' (digital audio tape) (Ibid: 87-90).

Using the metaphor I coined in concluding my comments on Renouf's 2007 work, it appears that Fischer studies either the trunk of the 'neologic tree' (the original neologism, for example 'cyborg', a blend of 'cybernetic organism' (although whether 'cybernetic' is considered new is not mentioned) (Ibid: 99-100)) or the branches of new derivations sprouting from previous neologisms, for example the clipping 'techno' giving rise to words such as 'technophobia' and 'technocracy' (Ibid: 148-153). The only reference to any derivatives of 'cyborg' is a brief mention of the abbreviation 'borg' and the further blend 'cyborganic' (Ibid: 100).

The Fischer book as a whole is dense and difficult to understand, without even an index to aid the reader in navigating the text. There is limited and inconsistent linking of the lexicographical and media elements of the study (something which my own research project seeks to address), with, for example the discussion of 'cyber' in the newspaper section merely noting which dictionaries the derivative forms came from, yet engaging in no further discussion (Ibid: 141-2). The methodology is unclear and although it initially appears that Fischer may be seeking to devise a theory of creative neologisms, there is no hypothesis or research question attached (Ibid: 1-3) and I can

find no further mention of this. Thus given the difficulties in accessing the information contained within Fischer's text, the age of the neologisms (up to 25 years old at the time of writing) and the degree to which *The Guardian* corpus has moved on, I will not be using this to inform the design of my own study.

2.3 Lexicography, Corpora and Social Media

As mentioned in 2.1, studies of 'neologism + lexicography' are few and far between, and this scarcity becomes even more pronounced when we move into the realm of 'neologism + e-lexicography' (see 2.3.1), the context in which I examine new words in this study.

Since the focus of this research project is 'lexicographical explorations of neologisms in the digital age', in this section I confine my review of the literature to that centring on dictionary making and new words, as well as key issues arising during the comparison of dictionary representations of neologisms, in particular standard and non-standard components of dictionary entries (see 3.4.3).

I therefore review a number of studies dealing with lexicography (traditional and, more pertinently, electronic), as well as corpora as tools for the creation of dictionaries, and the impact of social media on the making of dictionaries in the digital age. Specific topics covered include comparisons between expert-produced and collaborative dictionaries (including the role of contributors in collaborative dictionaries and the use of standard/non-standard dictionary entry components), and the relationship between corpora (as tools of dictionary making) and neologisms, particularly web-based corpora. It is perhaps useful at this point to reiterate the fact that the current research project is not a corpus linguistics study, but a lexicographical one, and therefore corpora are of interest only in their role as aides to dictionary-making, or in relation to neologisms.

The key studies to be critiqued in this section are: Abel and Meyer (2013), Meyer and Gurevych (2012), Atkins and Rundell (2008), Penta (2011), Moon (2009), Kilgarriff

(2013) and Grefenstette (2002). Of course a number of other sources will also be touched upon in addition to these.

2.3.1 (E)-Lexicography and Wiktionary

Although Algeo states that 'the history of English lexicography begins with the study of neology' (1993: 281), studies specifically on neologisms and lexicography are, as noted in 2.1, surprisingly lacking, and it is this gap which I set out to begin to fill with my own research project.

In the absence of more comprehensive studies to critique here, I turn instead to issues of e-lexicography and collaborative dictionaries, and in particular comparisons between expert-produced and collaborative dictionaries. Most notable of the issues here are the role of the collaborative contributor, and the use of dictionary components to represent new words, both of which form major elements of my own research. While traditional dictionaries have been in existence for hundreds of years, collaborative dictionaries have only been around for 20 to 30 years, with research into the area still in its early stages.

2.3.2 Corpora in Dictionary-Making

In this section, I briefly outline the use of corpus linguistics as a lexicographical tool, in particular the use of web-based corpora as a source of lexicographical data. For the past 30 years, new expert-produced dictionaries have been created using corpora. The addition of *Wiktionary* and other similar collaborative works means that we now have three different types of dictionary creation format:

- corpus-based, which are largely promoted as being based on a corpus, such as the *Oxford English Corpus* (*OEC*), for example the *Oxford Dictionary of English*
- corpus-informed, which are promoted as being created mainly from Reading Programmes and Citations, such as the *Oxford English Dictionary*
- Collaborative dictionaries, like *Wiktionary*, with no corpus involvement at all.

This distinction is not something which previous researchers investigating dictionary formats have, to my knowledge, considered in direct comparison, although there are

studies which touch upon some of the issues, for example those investigating whether corpus-based or lexicographer-written dictionary elements (usually examples) are more effective for users. One such study is Laufer's 2008 comparison of examples. Other authors compare expert-produced dictionary entries with those in collaborative dictionaries like *Wiktionary*, but although they may make reference to corpora, it is not as part of their comparison process (see for example Penta 2011, Meyer and Gurevych 2012). None of the comparative articles I have found has broken dictionaries down into all of their corpus-related formats and reviewed them on this basis. Indeed Laufer's 'corpus-oriented' classification is the closest I have found to open examination of lexicography influenced by but not fully dependent upon corpora (2008). Hence another gap which this study seeks to begin to fill is the lack of any literature on the relationship between 'corpus-based', 'corpus-informed' and 'collaborative' dictionaries.

It is perhaps useful at this point to distinguish certain terms. Hanks uses the accepted terms 'corpus-based' and 'corpus-driven' when discussing the COBUILD project (see below) (2012: 62). He does not define these terms, however. My interpretation is that 'corpus-based' means the information inside the dictionary comes from a corpus, while 'corpus-driven' indicates that the design of the dictionary and all of the processes used to create it were influenced by the corpus. I also employ the term 'corpus-based' (which I define in the same way as my interpretation of Hanks' usage), but this is used in juxtaposition with 'corpus-informed' rather than 'corpus-driven'. For the purposes of this study, 'corpus-informed' relates to dictionaries which obtain most (although not all) of their data not from corpora but from citations and Reading Programs. (It should be noted that I use these terms in a broader sense than Tognini-Bonelli (2001)).

The use of corpora as a tool in dictionary making began in earnest in the 1980s with the *COBUILD* project (a joint venture between the University of Birmingham and Collins Publishers) (Moon 2009: 436). This led to the publication of the *Collins COBUILD English Language Dictionary* in 1987, claimed to be the first corpus-based dictionary of English (Ibid). Corpora were not new at this time; the one-million-word Brown corpus had been around since the early 1960s (Kilgarriff and Grefenstette

2008: 90), but this was the first time that corpus linguistics had been used in a dictionary-making context (Moon 2009: 436). The COBUILD project adopted a new approach to several key lexicographical elements, including examples and defining styles (discussed in Section 3.4.3), however as will be shown, not all of these innovations have been considered successful in the long term. The original *COBUILD* corpus held 7.3million words (Sinclair 1987: 150) (compared to *Brown*'s one million), although this had grown to 18 million words by the time the first COBUILD edition was published (Hanks 2012: 62). This was followed in the early 1990s by the *British National Corpus*, which contained 100 million words, 90% of which were from written texts (Grefenstette 2002: 200-1). By 2012 the *Oxford English Corpus* (*OEC*)[8] contained two billion words.

The *OEC* is one of the key data sources for the *Oxford Dictionary of English* (Stevenson, 2010: ix) which Hanks claims to be (at the time of writing) the only corpus-based monolingual English dictionary aimed at native speakers (COBUILD being a learner dictionary) (2012: 62). Hanks goes on to examine corpora and lexicography through the lens of two specific approaches (Frame Semantics and Corpus Pattern Analysis), neither of which is relevant to the current study (Ibid: 65-76). He then takes a brief look at *Wiktionary*, but from the perspective of improving it through the use of corpus evidence (Ibid: 77-82). This lies in opposition to my own approach of comparing the collaborative format with those influenced by corpora, and hence I do not review it here.

One direction in which Hanks unfortunately does not look is that of corpora being built from web-based data, such as the *OEC*. With the continuing explosion in technology, this potentially represents the next big change for dictionary-makers, and while it has been a possibility for some years now, I have found only two papers discussing the issue: Kilgarriff 2013 and Grefenstette 2002.

Clearly Grefenstette's paper is somewhat dated now, and the technology prompting and populating it has in many cases been superseded (for example the webcrawling software currently available), however the underlying content is still valuable. He

[8] https://en.oxforddictionaries.com/explore/oxford-english-corpus

starts from a similar point as the researchers mentioned above, highlighting the opportunities presented by the theoretically endless space available to lexicographers producing information for electronic dictionaries. He quickly moves from this to the potential benefits of using the web as a corpus and seeing 'how this will change how lexicographers model word meaning' (2002: 199). Writing in 2002, the largest corpus he makes reference to is the *British National Corpus* mentioned above (100 million words), but he goes further, claiming that much of the 90 million written words (the other 10 million being spoken) are actually taken from newspaper articles (Ibid: 200-1). This not only lends credence to my own choice (in this project) of newspapers as the medium in which to monitor neologism use, it also provides a contrast to texts drawn directly from the web, some of which may not have undergone such rigorous editing and reviewing as a newspaper article, instead being 'dirty', or containing errors including spelling and grammar (Ibid: 201).

Kilgarriff refers to *UKWaC*, a corpus of British English drawn from the web and limited only by the .uk domain name (2013: 78). Built by 'webcrawling' it contains more than two billion words and was believed to be 'the only web-derived, freely available English resource with linguistic annotation' (Ferraresi, Zanchetta, Baroni and Bernardini 2008: 1). The EnTenTen12 web-based corpus (compiled in 2012 and available through Sketch Engine[9]) holds a little under 12 billion words, demonstrating at what pace such corpora are growing (Jakubíček et al 2013: 125).

Both Kilgarriff and Grefenstette review the tools available to lexicographers working with web-based corpus data, although Kilgarriff points out that those he mentions are part of the Sketch Engine software developed by himself and colleagues (Kilgarriff 2013: 78, 79-96; Grefenstette 2002). His paper is by far the more detailed of the two, covering issues such as lemmatisation (finding the base form, or lemma, which comprises the dictionary 'headword') and dictionary labels such as register, domain, region and those containing grammatical information (Kilgarriff 2013: 79, 87-90). He also covers examples and the 'GDEX' (good dictionary examples) algorithm in Sketch Engine which automatically selects the sentences most appropriate for use as

[9] https://www.sketchengine.co.uk/

examples in a dictionary (Ibid: 91). Although Kilgarriff's paper focusses on analysis of corpus data, in reality almost all of the processes he mentions could apply equally to non-web-based corpora. Indeed I think it is the size of the corpora in the article that leads to the idea that they are web-based, since the web is generally the source of such large corpora. Indeed in the published version of the chapter (see References) the title reads *Using Corpora as Data Sources for Dictionaries* and in the version downloaded from the Sketch Engine website (webpage since removed) the title reads *Using Corpora (and the Web) as Data Sources for Dictionaries*. This suggests either some vacillation about the title on the part of the author, or more likely an executive decision on the part of the book editor.

Grefenstette, meanwhile, remains clearly focused on web-based corpora, initially centring his discussion on collecting data, and then extracting 'recurrent patterns' (2002: 199-209). This begins with tools such as webcrawlers and search engines, then moves on to processes such as parsing (analysing the component parts of a sentence), tokenisation (identifying individual tokens or instances of specific words) (Kerremans, Stegmayr and Schmid 2012: 74) and KWIC analysis (key word in context) (Grefenstette 2002: 207-211). Kilgarriff (but not Grefenstette) mentions neologisms, referring directly to web-based data and commenting that the accepted way to identify new words is to use a 'monitor' corpus, comparing a current corpus with an older baseline corpus and extracting words in the former but not the latter as possible neologisms (Kilgarriff 2013: 81-3). He states that potential neologisms must have appeared 'in at least three or four documents' in order to be included in the neologism 'candidate list' (Ibid: 81), but this is slightly confusing since corpus-based dictionaries like the *Oxford Dictionary of English* (*ODE*) require more evidence of use before they accept a word[10] (see 3.4.2). Three or four documents is closer to the inclusion criteria for collaborative dictionary *Wiktionary*[11]. Perhaps there is a subsequent process where further evidence of neologism use is monitored until it reaches the level required for entry in a corpus-based dictionary. This may have been conducted through the *Oxford English*

[10] https://www.oxforddictionaries.com/news-and-press/oxford-dictionaries-faq
[11] http://en.wiktionary.org/wiki/Wiktionary:Criteria_for_inclusion

Corpus for the *ODE*, or it may be that this is the role of the new *New Words Corpus*, for which little information is currently available[12].

Both authors mention the use of 'word sketches' as a form of analysis (Ibid: 83-6; Grefenstette 2002: 211-214). Kilgarriff defines a word sketch as 'a one-page, corpus-based summary of a word's grammatical and collocational behaviour' (2013: 83). However these can be drawn from any corpus containing sufficient data; they are not limited to web-based corpora. Conversely a small web-based corpus may not contain enough data to generate word sketches. Although the database created and discussed in this research project contains 4.2million words, there are a number of neologisms where there is insufficient data from this web-based database for Sketch Engine to be able to build word sketches.

2.3.3 The Impact of Social Media on Dictionary-Making

To my knowledge, there are no specific sources addressing the impact of social media on lexicography, and therefore here I draw together a number of articles on the management of information in a digital age, as well as on the people who choose to actively engage in this new online society.

Collaborative dictionaries are one of a growing number of similar resources formed as part of the explosion in social media (defined as 'the set of tools that "enable people to connect, communicate and collaborate" … [and] include blogs, "wikis", social network sites') (Hemsley and Mason 2012: 3928). This follows the introduction of interactive Web 2.0 technology (discussed in 1.1). The most renowned of these is, of course, *Wikipedia*[13], launched in 2001 and claimed to be 'among the most prolific collaborative authoring projects ever sustained in an online environment' (Bryant, Forte and Bruckman 2005: 1). *Wiktionary* was created a year later, and while Bryant, Forte and Bruckman's article deals with contributors to *Wikipedia* (known as *'Wikipedians')*, the similarities between the two sites suggests that a similar analysis could be conducted on *Wiktionary* contributors.

[12] http://www.oxforddictionaries.com/words/oxford-new-words-corpus
[13] https://en.wikipedia.org/wiki/Main_Page

Social media sites like these have been shown to link individuals and groups across time, geography and culture, allowing them to share knowledge and ideas in a way that was never possible before (Bryer 2013: 45). Certainly the size and breadth of *Wiktionary* is something new in the field of dictionaries, offering a previously unheard of degree of involvement and collaboration in the field. It would be interesting to see an article like Bryant, Forte and Bruckman's investigating in similar depth what it means to be a *Wiktionary* contributor. Their descriptive study explores the experience of contributing to *Wikipedia* in the context of social activity, specifically Activity Theory and Legitimate Peripheral Participation:

> Activity Theory suggests a structure for thinking through technology use and emergent social norms on Wikipedia and how they influence the transformation of members' participation over time (Bryant, Forte and Bruckman 2005: 3)

Legitimate Peripheral Participation meanwhile refers to how new members of a group gain acceptance by first completing 'peripheral tasks' that benefit the whole community (Ibid: 2). We could perhaps call the increased involvement in dictionaries and the language at large a form of 'linguistic citizenship'. Bryer (2013) has shown how social media can be used to encourage active citizenship, enticing and engaging citizens not only to keep them informed but also to highlight their rights and responsibilities and provide them with a space in which to discharge them.

Hemsley and Mason discuss the changes social media is bringing to the way knowledge is managed and shared, in particular the use of 'wiki' sites (defined by Leuf and Cunningham as 'a freely expandable collection of interlinked web "pages"' (2001: 14)) where knowledge is considered 'not 'static', but rather ever-changing and immediate' (Hemsley and Mason 2012: 3929). They conclude that 'the widespread use of social media creates a dynamic, recursive socio-technical information and knowledge sharing system'. This system is also known as a knowledge ecosystem, which like any other ecosystem must require care and management to survive (Ibid: 3928). The fact that all of these changes and all of this information comes from ordinary users leads to the idea of collaborative sites as a form of 'crowdsourcing'.

However the same term also can be applied to 'historical' dictionaries such as the *Oxford English Dictionary*, which have long used external 'readers' to collect examples illustrating how words are used. Currently the same methods are being used to gather information on World Englishes, or lesser known varieties of English from across the globe[14].

2.3.4 Wiktionary *and Other Collaborative Dictionaries*

A number of papers have been written in recent years on the rise of collaborative dictionaries, examining issues such as word senses, the impact on the wider field of lexicography, and user contributions (see for example Meyer and Gurevych (2010); Penta (2011); Abel and Meyer (2013)). Abel and Meyer, for instance, focus on user contributions, providing a useful 'roadmap' of scholarly interest in the development of collaborative dictionaries to date, before moving on to propose a 'functional classification system for user contributions to online dictionaries' (2013: 180).

Abel and Meyer's review of literature on the topic is revealing, demonstrating how views of collaborative dictionaries have developed over a 15-year period, from Carr's original proposition of 'bottom-up lexicography' (in which dictionaries 'evolve upward from readers' (Ibid: 181; Carr 1997: 214)) through Køhler Simonsen's description of the evolution of lexicographic services (Abel and Meyer, 2013: 181; Køhler Simonsen, 2005) and Storrer's (2010) comparison of true collaborative works with traditional dictionaries inviting user contributions (which I have been unable to find in anything other than the original German, which I do not speak) (Abel and Meyer, 2013: 181). Finally Abel and Meyer draw attention to Lew's discussion of degrees of user-generated content across dictionary types (Abel and Meyer 2013: 182; Lew 2013).

Several of these works now feel very outdated, with the realities of technology and social media having outstripped the work on which they were based (see for example Carr 1997; Køhler Simonsen 2005). Nevertheless, they provide a useful background against which to paint future research, as does Lew's 2013 work, in which he not only takes a detailed look at collaborative dictionaries like *Wiktionary* and the *Urban*

[14] See http://blog.oxforddictionaries.com/2014/02/can-world-englishes-benefit-crowdsourcing/

Dictionary, but also examines user-generated content in 'professional' works, particularly *Macmillan's Open Dictionary*. He calls these 'collaborative-institutional dictionaries' (Lew 2013: 6-12, 15-17; Lew 2011: 237). I term them 'semi-wiki' dictionaries, and there is danger of confusion here, since Melchior (2012: 1), uses a similar term ('semi-collaborative') to refer to a slightly different format, that of dictionary portals like *LEO*[15]. Here, users can make suggestions or discuss issues in forums, but there is no over-arching editorial control (Nesi 2012: 374). What is clear is that, as with the phrases in 2.2.1, care must be taken in defining terms in order that we not become confused by other writers' framing of the dictionary landscape.

While Abel and Meyer's paper provides useful background on the rise of collaborative dictionaries, there are a number of issues. Placing all those who make any kind of contribution to a dictionary in the same category is, in my view, overly simplistic. Based upon my reading of Discussion, Talk and Profile pages for *Wiktionary* contributors, I would have to say that many of them appear to consider themselves more than simple 'users' of the site, but actual 'partners' in its success (although this partnership is heavily influenced by the templates and guidelines used to try and ensure consistency across *Wiktionary* entries[16]). In my opinion these contributors have more freedom and the opportunity for more involvement than those on other collaborative dictionary sites, for example the *Urban Dictionary* (2013), where information can simply be added, either to existing entries, or as discrete new entries[17]. In addition, some of the actions which Abel and Meyer put forward as being user contributions are actually just user personalisation, for example the selection of favourite articles in *Dictionary.com* (2013: 188). Individuals from whom 'log file analysis' feedback is gained, meanwhile, must, in my view, be considered much more passive (Abel and Meyer 2013: 182). In addition the extent of the changes that can be made by *Wiktionary* contributors is not made clear by Abel and Meyer. These issues are better addressed in Meyer's and Gurevych's 2012 chapter, as will be discussed below.

[15] http://dict.leo.org/
[16] https://en.*Wiktionary*.org/wiki/*Wiktionary*:Entry_layout
[17] http://www.urbandictionary.com/

Meyer and Gurevych (2012) offer the most detailed overview of *Wiktionary* and how it operates, as well as comparing it with other online lexicographical resources. They state that their chapter explores 'the possibilities of collaborative lexicography' through study of *Wiktionary*, which they claim is 'the largest available collaboratively constructed lexicon for linguistic knowledge' (2012: 260).

The chapter presents the history of how *Wiktionary* was created, along with the many language versions it comprises. The authors describe the structure of the site, the opportunities for collaboration and the guidance provided for these processes (Ibid: 260-274). While none of the other articles on *Wiktionary* cover its structure in such detail, there are several issues here that I disagree with. As any user of *Wiktionary* can discover, the Talk pages attached to entries (part of the collaborative infrastructure of the site) are not accessible as suggested by Meyer and Gurevych (2012: 272). Instead, a *different* version of the page is accessed through a kind of 'back door' from the search page for the 'Tea Room' (the main discussion forum). However the information contained within the two is not the same (it is not clear why). In addition, although Meyer and Gurevych state that each entry's Talk page 'can be used to discuss its content' (Ibid), in fact these pages are clearly marked with instructions from administrators that they should not be used because they are not regularly reviewed by other users, and hence a debate is unlikely to take place there.

Meyer and Gurevych then move on to critique *Wiktionary*. It is noticeable, however, that while they discuss many aspects of dictionary entries, both from an informational and a critical standpoint (Ibid: 260-274, 274-89), even noting that *Wiktionary* 'has no fixed structure for its entries' (Ibid: 268) they make no mention of the standardised, industry-accepted components which make up a traditional dictionary entry. Indeed the structure of Meyer and Gurevych's chapter makes it difficult to even try and map these standardised components onto the elements of a *Wiktionary* entry since in many cases they approach the issue from a more conceptual level, for example talking about 'semantic knowledge' rather than breaking this down into its component parts such as definitions and examples (Ibid: 270). The standardised components are explained by Atkins and Rundell in their *Oxford Guide to Practical Lexicography*, in which they provide detailed information on all aspects of a traditional dictionary entry (2008: 202-

246, 385-462). Among these are the differing defining styles that can be used to present the meaning of a word in a dictionary, of which Meyer and Gurevych make no mention. I discuss these issues in light of the current project in 3.4.3.

Throughout the chapter, Meyer and Gurevych compare the English *Wiktionary* with its German and Russian counterparts, and with multiple 'expert-built lexicons' in each language. As mentioned in 2.2.3, these 'expert-built lexicons' are not the same as the 'expert-produced dictionaries' used in my own study and I hence do not review this section of the Meyer and Gurevych study since it would not be comparing 'like-with-like'.

In their paper there is a small section on neologisms in *Wiktionary*, however this relates only to numbers in the different language versions, and there is no discussion of where the words come from, how they enter the dictionary or how their entries may develop over time (2012: 277).

However while Meyer and Gurevych's 2012 paper offers a more detailed picture of *Wiktionary*'s structure and functions, Penta's is the more critical. He reviews the current and historical state of lexicography, and compares a *Wiktionary* 'article' with its corresponding 'entry' in the *Oxford English Dictionary* (*OED*) and the *Urban Dictionary* (2011). Through this, he discusses (and in many cases rebuts) criticism of the collaborative model, for example arguing that professional lexicographers and amateur contributors are not as different as some might suggest (2011: 3-7). However this support for the collaborative form is somewhat undermined by his failure to note that amateur contributors may nevertheless be experts in the field on which they are contributing, something which Meyer and Gurevych do acknowledge (2012: 259). Reviewing contributors' profile pages on the site would have provided Penta with the knowledge needed to make this connection.

It is reassuring to note, however, that Penta recognises the rigorous nature of *Wiktionary*'s inclusion criteria, and the fact that although its pages were originally populated from copyright-expired volumes such as *Webster's New International Dictionary of the English Language* (Meyer and Gurevych 2012: 262) it is no more

guilty of plagiarism than most traditional dictionaries (Penta 2011: 4). While his discussion of these wider issues of lexicography is of interest, his comparison between dictionary types is limited, focusing on a single headword ('bomb'). However even bearing this in mind, it is more useful in setting the scene for my own research than Meyer and Gurevych's paper, which compares the English version of *Wiktionary* with Princeton University's lexical database *WordNet* and *Roget's Thesaurus* (2012: 274-291). Meyer and Gurevych conclude that *Wiktionary* does represent a credible rival to expert-produced dictionaries, for example providing information which is not included in traditional works (such as translations for numerous entries (2012: 280)), immediately updating to publish any changes (rather than waiting for scheduled updates), and including more neologisms than expert-produced dictionaries (Ibid: 277-8). However all this is for nought in terms of setting the scene for my research project, when we consider that *WordNet* and *Roget's Thesaurus* are databases rather than dictionaries and hence do not represent a like-with-like comparison (as discussed in 2.2.3).

Penta concludes that, with regard to the presentation of different meanings for the word 'bomb', 'cyber-lexicons are on par with the *OED* in handling semantic information' (2011: 10), and that with regard to the tools provided to 'assist the reader' in decoding definitions ('illustrative examples, usage notes and hyperlinked text'), the collaborative dictionaries actually outperform *OED* (Ibid: 10-12). Of course the possibilities offered by electronic formats such as collaborative dictionaries exist only as a result of the enormous changes experienced in the field of lexicography, as it has morphed into 'e-lexicography' over the past 20 to 30 years. Indeed according to Penta (2011: 2-3) and Lew (2012: 344, 361), tension still exists around the question of whether emerging technology should be viewed as offering a whole new platform upon which to redesign the very idea of what makes a 'dictionary' (a 'clean slate'), or whether it simply presents an opportunity to speed up and streamline existing functions. Indeed Penta claims that 'we are witnessing a paradigmatic shift of authority in which users, rather than editorial boards, are making decisions concerning the content associated with a lexical entry's definition' (2011: 1). In my opinion, and that of Chen (2013), the potential offered by the 'clean slate' approach is too great to

be ignored. However others appear less certain, with Lew seeming to vacillate between the two viewpoints. However as his final comment on the matter seems to suggest that electronic dictionaries can forge ahead provided their designers break free of traditional thinking, we can assume that he, too, favours the more radical approach (Lew 2012: 343-4, 361).

2.4 Automated Systems for Collection of Web-Based Corpus Data: The *NeoCrawler*

One of the most recent studies using automated systems to identify, monitor and analyse new words, is that of Kerremans, whose PhD thesis (published in 2015) on '*A Corpus-Based Study of the Conventionalization Process of English Neologisms*' includes the creation and use of the *NeoCrawler* program. This was developed by a team (including herself) at Ludwig Maximilian's University in Munich. An additional article was published by Kerremans and her colleagues Stegmayr and Schmid the same year as she submitted her thesis (2012). As a result of having presumably been written several months earlier, the latter contains a number of inaccuracies/elements which in the thesis were either changed or reassessed (see below). In this section, I review and evaluate Kerremans' study and the team's supporting article, particularly with a view to the comparison made during the course of the current project between this automated system and my new manual methods of corpus data collection.

The *NeoCrawler* was an automated webcrawler and linguistic analysis system which collected data from the *Google Blogs* environment between 2006 and 2011. (Sadly the *NeoCrawler* ceased operations shortly after this, due in part to lack of funding (Personal Communication, Kerremans 2013). I have been able to find no results in the system beyond early 2012, indicating that it remains inactive.) Kerremans' project set out to explore the 'conventionalization' process of new words, 'conventionalization' being 'the dynamic socio-pragmatic process by means of which lexical innovation becomes established in the language and the speech community' (2015: 22). She specifically investigated 'alleged conventionalization-promoting and -inhibiting factors' by 'closely monitoring the social and linguistic diffusion behaviour of 44

English neologisms in the online speech community' (Ibid: 227). She also sought to investigate the 'emergence of syntagmatic lexical networks during the conventionalization process, again based on longitudinal data retrieved from the Internet' (Ibid). The Kerremans, Stegmayr and Schmid article had similar objectives, aiming to answer questions about why some new words are successful in becoming established, and others are not. They then present the *NeoCrawler* as a tool for addressing such questions with regard to online data (2012: 59).

To gather the data for her doctoral study, Kerremans used the *NeoCrawler* system, which identified first coinages of new words, tracked their use online, and provided tools for socio-pragmatic analysis of the resulting neologisms (2015: 25, 78-84, 84-92). The *NeoCrawler* comprised two key components: the *Discoverer* which searched for first instances of neologisms, and the *Observer* which then tracked their development and conducted the socio-pragmatic analyses of their behaviour (Kerremans 2015: 78, 84).

Whilst a very comprehensive study, Kerremans' thesis is at times somewhat dense and confusing. (Although this may be simply due to the fact that English is not her first language and the thesis was written to meet the requirements of a different academic tradition (German)). Part of the problem is that there is no discrete Literature Review, and hence her evaluations of other authors' works are spread throughout her work, making it often difficult to identify whether a thought or action is her own, or something she is reviewing. It may be that this is standard practice in Germany, but it does add unnecessary complications to reading a thesis or book. One thing which is never made clear is which area of linguistics the study is intended to inhabit. Clearly there is an element of corpus linguistics, since the *NeoCrawler* is designed to work with corpora. It seems that the thesis also falls into the field of computational linguistics, due to the application of automated techniques not only to the collection of data, but also the programming of the *Discoverer* by a Computational Linguistics postgraduate student (Kerremans, 2015: 80). However there is also the socio-pragmatic analysis of neologisms to consider, suggesting a third field of study. None are ever overtly stated. Kerremans makes brief mention of neologisms and lexicography, but this seems to be simply because of the integral relationship between

new words and dictionaries (Ibid: 17-18), and not because of any attempt to position the work in this field. This and the analysis of newspaper articles rather than blogs are the two areas in which my own study is designed to expand upon the work of the *NeoCrawler*, whilst at the same time comparing its automated methods with my own manual methodology.

The 2012 chapter on the same *NeoCrawler* study, by Kerremans, Stegmayr and Schmid is even more challenging to understand than Kerremans' thesis (2012). Even the case study 'detweet' is problematic. While it illustrates processes of lexicalisation ('changes in form and meaning') and diffusion ('increase in frequency of usage'), it is not clear which of the meanings mentioned is becoming institutionalised: the act of *not* tweeting, 'the removal of Twitter messages or tweets' or 'retweeting' with a tone of disapproval (Ibid: 60, 81-93). This complication is not adequately addressed in my opinion, with the authors merely concluding that 'so far *detweet* has only been institutionalised somewhat tentatively, because it has not started to disperse into more formal registers and text types' (Ibid: 90). Further confusion is caused in Kerremans' 2015 work through her discussion of the different ways in which new words can be located and tracked online, and this is also found in the wider literature on the issue of web-based corpora (as opposed to that specifically intended for lexicographical purposes). Some authors (such as Fletcher 2013) use the terminology web-as-corpus (WAC) and web-for-corpus (WFC). With WAC, the entire web is used as the corpus, accessed via commercially available search engines with corpus query tools attached, whereas with WFC the corpus is selected from the available material by 'webcrawlers', and downloaded for analysis (Ibid). Other researchers refer to webcrawlers in both contexts, distinguishing instead between 'on-demand' crawlers (believed to be the same programs as used in WAC) and 'downloadable' crawlers' (roughly equivalent to WFC). Kerremans (2015) and Kerremans, Stegmayr and Schmid (2012) do the same, causing considerable confusion whenever 'webcrawling' is mentioned, as to whether the web is being treated as a corpus or as a source of one.

A corpus is usually constructed using one method or the other, however in the case of the *NeoCrawler*, the *Discoverer* searched the web (much as do standard WFC programs), downloading pages containing potential neologisms identified in part by

their absence from the *NeoCrawler's* internal 'dictionary' (Kerremans 2015: 80-1). The *Observer*, however, operated as a kind of cross between a web-for-corpus (WFC) crawler and a web-as-corpus (WAC) search program, using Google to locate webpages containing neologisms identified by the *Discoverer* (WAC), but marking them for download (WFC) (Ibid: 84-6; Kerremans, Stegmayr and Schmid 2012: 62-65). This appears to fit the new model being developed by Renouf and Kehoe, which they call 'web as corpus shop'. In their model a 'tailor-made search engine' (a term which applies to both the *NeoCrawler* and Renouf and Kehoe's WebCorp LSE) 'is designed to download and process web texts for inclusion in structured, linguistically-analysed off-line corpora' (2013: 168). The use of the *NeoCrawler's* internal 'dictionary' to help identify potential neologisms (Kerremans 2015: 80-1) is another problem area. Several of these candidate words were, in my opinion, wrongly identified as 'new' by the *NeoCrawler*, as although they passed the program's internal dictionary check they were, in my view, already established. This leads to questions about the reliability of the processes used both to identify new words, and to determine when they first entered the lexicon.

During the testing phase, the *NeoCrawler* project was confined to the *Google Blogs* environment, and this meant that, whilst it might not have been the intention for later uses of the program, in this case the corpus created and analysed by Kerremans can be considered genre-specific, since it is limited to blogs only. This is something which I suspect she had not considered; she makes several references to different genres (see for example 2015: 69, 73, 78) and even puts forward a classification system including 'blogs', without ever seeming to recognise that she has, in fact, produced a 'genre-specific' (or to use her preferred term, 'type of source'-specific) corpus herself (Ibid: 90). However it may simply be that she was planning future phases of the project's development, since it is stated that the next step was to include other blog providers (Ibid: 80).

One element of 'post-processing' of the *NeoCrawler* data with which I disagree in both publications under review here is the claim that duplicate files can be removed from a list of Google search results by comparing the 'title and the file size to all previous results of the same search' (Kerremans, Stegmayr and Schmid, 2012: 73; Kerremans

2015: 70). I find that when small changes have been made to a newspaper article (usually adding a short reader comment, or making a minor amendment to the article text), the file size is unaffected. Provided the title has not changed, comparing current search results with previous ones does not, in fact, accurately reveal duplications. The researcher might therefore remove a newer article assuming it to be a duplicate when actually it contains new material.

One proposed development to the *NeoCrawler* system which Kerremans, Stegmayr and Schmid suggest (2012: 77) but that Kerremans does not, is that further automation was planned for the classification of texts and neologisms. This would, in my view, open the door for the introduction of additional errors (such as those caused by automated dictionary checking (Ibid: 81)), allowing words which should have been included in the dictionary to be marked as not-present. This is another area where there is some confusion between the two texts since the Kerremans, Stegmayr and Schmid chapter does not make it clear that the 'reference [internal] dictionary' mentioned above, which was used to check potential neologisms (2012: 80) was actually 'compiled from the English version of *Wikipedia* and a catalogue of N-grams representing known words' (Kerremans 2015: 81). Indeed the former lists the 'reference dictionary' as being in addition to the 'user-generated catalogue of known words' (N-grams). Similarly 'Google's University Research Program for Google Search' onto which Kerremans, Stegmayr and Schmid say that the *NeoCrawler* project had been accepted (2012: 67) was actually discontinued by the time Kerremans submitted her thesis (2012: 70).

Replicability will be a problem for the *NeoCrawler*, since to replicate a study, one must be able to exactly repeat it, copying all aspects of the methodology and achieving the same results (Lüdeling, Evert and Baroni, 2007: 10-11). This is not possible for the data collection phase of a web-based research project since webpages are constantly being removed, amended and re-uploaded, meaning that at any time a key piece of data could be unobtainable. It is likely that reproducing the study would be equally problematic, since finding webpages that exactly matched the characteristics of those in the original one would be almost impossible (Ibid: 11).

One of the most important elements missing from Kerremans' study is that of wider 'contextual' information, such as publication date or author. When discussing 'contextual' information, I base my own definition upon Sense 1 of the *Oxford Dictionaries* online (*ODO*) entry, adjusting it to read 'relating to the surrounding linguistic and extralinguistic information that forms the setting within which a neologism appearing in a newspaper article sits'[18]. Kerremans focuses heavily on 'cotext' (2012: 29-36), which in fact appears to match *ODO*'s second sense for the word 'context' as at 25 September 2016 (it has since changed): 'the parts of something written or spoken that immediately precede and follow a word or passage and clarify its meaning'. However she pays little attention to the kind of contextual information that I consider to be crucial when collecting data for a genre-specific corpus: information such as publication date and where on the page the neologism appears.

2.5 Conclusion

Dictionaries are crucial to understanding the world in which we live, for as Mugglestone points out, they 'profoundly engage with how we, as speakers, understand the world and articulate the nature of what we perceive' (2011: 16). As we have seen, neologisms keep that engagement current, naming and describing new words and concepts (often scientific or IT-related) (Lehrer 2003: 371; Francl 2011: 417; Mitchell 2008: 33)), often through the innovativeness of journalists (Renouf 2007: 70) and other professional writers who create these new terms (Francl 2011). For many consumers of language, their first exposure to new words is through the media, either by the press using existing new word/word formations, or actually creating their own (Fischer 1998: 68-9). While separate literature exists on both these topics (lexicography and neology), there is, as yet, very little written on how they interact, and as discussed, it is here that the current research project seeks to fill a gap in the academic research. It explores the relationship between neologisms as agents of language change and the dictionaries in which they ultimately appear. At the same

[18] https://en.oxforddictionaries.com/definition/contextual

time it draws a picture of these neologisms in the news media, and compares manual methods of corpus data collection with the automated one most recently written up, by Kerremans and her colleagues (2015 and 2012).

Gaps in the academic literature were also identified regarding differences between dictionary formats ('corpus-based', 'corpus-informed' and 'collaborative'), along with the lack of writing on methodologies for working with neologisms.

In this chapter a review of the academic landscape into which this study enters was provided through discussion of developments in (e)-lexicography and specifically *Wiktionary*, as well as the effect of social media on dictionary-making, and the importance of corpora as a lexicographical tool. Finally a detailed review of the *NeoCrawler* study was provided, as this forms the basis of this project's comparison between manual and automated methods of data collection, the findings from which are discussed in Chapter 5.

Chapter 3 Methods and Methodology

Part 1 – Laying the Groundwork

3.1 Introduction

In this chapter I present key elements underpinning this research project, including dictionaries used for comparing representations of neologisms, newspapers used to track usage of new words and web-based corpus data of the type which the new methodology devised here aims to collect. I also introduce a new program devised by Ludwig Maximilian's University in Munich, the *NeoCrawler* (see 2.4), which is the automated program against which the new methodology was compared. This chapter further lays out the methodological framework for the study, and discusses the importance of research validity and reliability, and how they can be achieved.

As outlined in 1.1, this study examined entries for a set of 34 neologisms in four expert-produced English dictionaries and compared them with corresponding entries in collaborative dictionary *Wiktionary.* To provide a complementary picture of real-word usage of these words, it also tracked these new words in online versions of national newspapers between 2000 and 2014. Due to the need for detailed contextual data in order to closely target the selection of newspaper articles, a new methodology was devised for the collection of web-based corpus data.

The purpose of the study was two-fold:

1. To compare degrees of comprehensiveness in the entries provided for new words in expert-produced dictionaries with those in collaborative dictionary *Wiktionary*

2. To track neologism appearances in UK news media in order to compare usage and behaviour in different newspapers at different stages in the neologic life-cycle

The second of these objectives depended upon the design, development and implementation of a new methodology of data collection aimed at creating context-

rich genre-specific corpora, of the kind used in lexicographical research. Corpus linguistics has a history of use within the field of lexicography, dating back to the beginning of the COBUILD project in 1980. Indeed Hanks states that 'the first major impact of corpora on lexicography was on a dictionary for foreign learners, namely COBUILD' (2012: 62; Renouf 1987: 1). This new methodology called for a further objective:

3. To consider whether neologism use and behaviour in the media can be best explored through the use of new manual or existing automated corpus data collection techniques.

Due to the nature of this project as an exploratory study of neologisms in the digital age, the methodology devised here was used to create a database of neologism-containing texts (the *NTON* database (*Neologism Tracking in Online Newspapers*)) rather than a corpus. A corpus is, according to Sinclair 'a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research' (2004: 22). The *NTON* database can be distinguished from a corpus by the fact that it was not intended to 'represent a language or language variety' (Ibid), but was instead intended to provide a picture of neologisms in a specific context: newspapers. Thus not all of the features of a corpus would be required.

3.2 Methodological Framework

In this section I present the methodological framework within which this study is conducted, exploring key issues such as validity, reliability and replicability.

It is widely recognised within the fields of lexicography and neology that there is no one methodology which can be applied to these topics, to the exclusion of all others. Indeed a mixture of methodologies is often considered the most appropriate approach, with qualitative methods guided by interpretative methodology, and quantitative methods by positivist tools. Many of these methods are drawn from the

realms of education and the wider social sciences, since language is central to understanding learning and the culture in which we live (Dörnyei 2007: 21; Matthews 2003: 26-8).

3.2.1 Positivist, Interpretative and Mixed Methodologies

Here, I explore positivist, interpretative and mixed methodologies, and their role within lexicographical and neological research. Positivist methodologies employ quantitative methods of data collection which 'result primarily in numerical data which is then analysed primarily by statistical methods' (Dörnyei 2007: 24). Such methods are popular with computational linguists, and with corpus linguists dealing with large amounts of data on issues such as frequency or collocation (see for example Hunston 2002).

Dating back to the Ancient Greeks, positivism holds that 'all genuine knowledge is based on sense experience and can only be advanced by means of observation and experiment' (Cohen, Manion and Morrison 2000: 8). Positivists believe that all academic research should follow the same forms as scientific enquiry, with the emphasis on objectivity and conducting research 'from the outside' (Ibid: 8, 22, 35). Such quantitative methodologies have the advantage of offering precise measurements and reliable, replicable data by virtue of their systematic and rigorous approach (Dörnyei 2007: 34). However positivism is criticised for failing to 'capture the real meaning of social behaviour', for example not taking account of the importance of factors such as freedom, choice, and individualism (Sarantakos 1998 cited in Robson 2002: 23; Cohen, Manion and Morrison 2000: 17). While studies based entirely in the positivist tradition are not to be found in the literature relating to this research project, some quantitative methods are employed in many of those studies, in order to provide a statistical context for the interpretative results obtained using qualitative methods (see below).

Interpretative methodologies, in contrast to the positivist tradition, place emphasis on the individual and his/her understanding of the world (Cohen, Manion and Morrison 2000: 21-2). Studies using these methodologies tend to be small scale, non-

statistical projects, employing qualitative methods such as discourse analysis, interviews and participant observation to measure subjective understanding of an individual's experiences. Research methods and indeed research questions may change and develop over the course of a study, meaning that qualitative researchers enter a project with an open mind; their work is not structured around previously conceived ideas or hypotheses, but is instead allowed to develop organically (Cohen, Manion and Morrison 2000: 21-3, 35; Dörnyei 2007: 37-8). This is indeed the case with the current research project, in which a number of early suppositions and ideas were rejected during the development of the new methodology, before the final research questions outlined in 3.9 were established. This, then, is an example of one such iterative study.

While such qualitative methodologies provide 'insider' understanding of subjects and situations, along with an unparalleled level of flexibility that allows the researcher to follow emergent research threads, they are not without their critics. It is argued that they lack methodological rigour, and place heavy emphasis on the insights of the individual researcher, and that this can limit their effectiveness, as can the time and labour required to carry them out (Dörnyei 2007: 37-42). It appears however that this argument is often made by quantitative researchers who are perhaps unfamiliar with the workings of qualitative research (Ibid: 41). When the qualitative researcher adopts the methods and approaches outlined above, rigour is achieved.

As this demonstrates, the distinction between positivist and interpretative methodologies is not black and white, nor are they mutually exclusive. Many research projects, including this one, employ a mixture of the two traditions, adopting either quantitative or qualitative methods depending on which is best suited to the task at hand (Robson 2002: 43).

Much lexicographical research shadows the processes used by lexicographers themselves, employing quantitative methods to collect and collate word usage, and complementing this with qualitative assessment and interpretation of the meaning and validity of that word and its various different senses (see for example Atkins and Rundell 2008: 78-92). For the researcher, this is followed by comparison of different

definitions, and investigation of users or of other associated factors (see for example Penta (2011), Meyer and Gurevych (2010)).

The study of Neology (lying at the root of English lexicography, according to Algeo (1993: 281)) works in much the same way. Algeo himself used numerical analysis to show the break-down of word formation processes amongst new words following a qualitative discussion of where selected terms had come from (1980). Similarly, he uses quantitative methods to explore how many new words had survived since the closure of their corpus, before categorising them and engaging in a qualitative discussion of these categories, much as did Renouf (Algeo 1993; Renouf 2013).

3.3 Research Validity and Reliability

In the following sections I explain how issues of research validity and reliability are addressed in this study, particularly with regard to replicability, reproducibility and representativeness. These issues can be especially problematic in studies such as this, where data for a database or corpus is collected from the World Wide Web.

A crucial element of any piece of research (whether it be quantitative or qualitative) is the validity and reliability of the study, since this proves that the findings produced are robust and relevant and that they can be extrapolated to wider populations and contexts (Dörnyei 2007: 50). Central to the questions of 'validity' and 'reliability' are the issues of replicability, reproducibility and representativeness.

3.3.1 Replicability, Reproducibility and Representativeness

Replicability is the ability to exactly repeat the study, copying all aspects of the methodology, and achieving the same results (Robson 2002: 42). Doing so demonstrates the rigour, control and precision of the study (Dörnyei 2007: 34). An alternative approach which lends similar credence to the original study is to seek to reproduce the same results from different sources (in this case, different dictionaries and different databases or corpora) compiled in exactly the same way: reproducibility (Lüdeling, Evert and Baroni 2007: 17). While this would not be problematic for the

dictionary comparison portion of this study, one of the difficulties with replicability and reproducibility for web-based corpus research is the dynamic nature of the web itself. As Fletcher (2013: 1) points out, 'the Web is constantly expanding', citing Alpert and Hajaj's assertion that 'several billion (10^9) new Web pages appear daily' (2008). This enormous growth might not be such a problem, were it not for the way in which the constantly expanding data is presented to users. Search engines such as Google Advanced Search (GAS) (used in this study) do not simply add new data chronologically, but rather mingle results together, in response to companies' 'search engine optimisation' efforts. 'Search engine optimisation' is 'the process of trying to rank highly a given web page or domain for specific keywords' (Evans 2007: 22). These keywords can be placed in webpage titles/headings and metadata indicating an organisation's main areas of business. These, along with elements such as extensive use of 'in-links' (URLS [Universal Resource Locators] referencing specific web pages), are coded in such a way as to encourage search engines to place them near the top of any Search Results Page (SRP) (Ibid, citing Pringle, Allison and Dowe 1998; and Fortunato et al 2006). Thus rather than new results being easily identifiable due to their addition to the first few SRPs, they are instead inserted throughout them, depending on their perceived relevance to the queried search word.

The problem is further complicated by the fact that the 'indexing and search strategies' of commercially available search engines can be altered or updated without warning and at any time, meaning that repeating searches, such as those conducted here within the domains of specific UK national newspapers, generates SRPs that bear little or no resemblance to each other (Lüdeling, Evert and Baroni, 2007: 11).

This not only presents problems for future corpus linguistics researchers wishing to replicate or reproduce a study, but also means that it is impossible to return to the SRP for a particular neologism in a particular text (for example to gather additional contextual information) since the SRP can look completely different each time the search word is queried. This was the case here, even when the interval was just a few days, due to factors such as website maintenance (pages being temporarily taken down and reposted after changes to the site), adverts and preferred results being

inserted and replaced by Google (based upon a user's previous browsing history and location (Fletcher 2013: 3)) and various other algorithms used by search engines to provide as many search results as possible.

3.3.2 Representativeness

Also important to the validity and reliability of a research project is the representativeness of the data. In order to be useful, a corpus must represent both the language from which it is compiled, and the subset of language(s) under study (Sinclair 2004: 6). In this case that is English neologisms in UK national newspapers. 'Representativeness refers to the extent to which a sample includes the full range of variability in a population' (Biber 2008: 63), or the degree to which its findings can be generalised to apply to the language/subset of language as a whole (Kennedy 1998: 62). Often size is considered the best way of achieving representativeness within corpus linguistics, the assumption being that the more texts and the more words included in the corpus, the more likely it is to be representative. However, it is often unclear exactly what it is that the corpus should represent, and this is particularly so in the case of web-based corpora (Kilgarriff and Grefenstette 2008: 97, 99; Fletcher 2013: 3), although for the *NTON* database (*Neologism Tracking in Online Newspapers*), the situation is slightly clearer, since it is specifically journalistic writing that the data seeks to represent

More important to the achievement of a representative corpus than size is the construction of the corpus itself – for example a solid understanding of the purpose and goal of the corpus, the setting of clear criteria for identification and prioritisation of text types, and sampling techniques (Sinclair 2004: 10-13). These issues were much less complicated in the present study than for many web-based corpora, since this project sought to track the use of neologisms in specific UK national newspapers. There are a finite number of such news publications, and within those, a finite number of articles containing these new words, meaning there was no requirement for decisions to be made on which parts of the language 'population' (use of the language under study) (Biber 2008: 63) to include and which to exclude. The questions here, then, were:

- Which neologisms to include in the study (see 4.2)
- Which newspapers to include in the study (see 3.5)
- Which dictionaries to include in the study (see 3.4)

3.4 Elements of Project: Dictionaries

In this section I present the dictionaries selected for this study and outline the reasons behind this choice. I also explore the relationship between dictionaries and corpora, specifically 'corpus-based', 'corpus-informed' and 'collaborative' dictionaries, and introduce some of the other differences between *Wiktionary* and the expert-produced dictionaries used in the study, including some of the standard and non-standardised dictionary components discussed in section 3.4.3 (see Atkins and Rundell (2008) and Meyer and Gurevych (2012)).

Objective 1 of the current study was to compare degrees of comprehensiveness in the entries provided for new words in expert-produced dictionaries with those in collaborative dictionary *Wiktionary*. In the latter, users are invited to contribute by adding, editing and removing entries. These were compared with dictionaries produced by 'traditional', expert publishers such as Oxford University Press. Since one of the central themes of collaborative dictionaries is their ability to be updated on a regular basis (Ibid: 259), online dictionaries were chosen for comparison (although no expert-produced dictionary can be updated as regularly as collaborative dictionaries since the latter are updated every time a user makes a change to the site. Expert-produced dictionaries are generally updated on a quarterly basis (Weiner 2009: 401)). One exception to this decision was the inclusion of the 2010 printed version of the *Oxford Dictionary of English* (*ODE*). This was done to provide a starting point against which additions to the online version (*Oxford Dictionaries*) could be assessed. *Oxford Dictionaries* online (*ODO)* is the online version of *ODE* (according to the *ODE* Preface (Stevenson 2010: vii)). *ODE* was first published as the *New Oxford Dictionary of English* (NODE) in 1998. *NODE* was largely based on the *British National Corpus*[19] which

[19] http://www.natcorp.ox.ac.uk/

contained 100 million words (Stevenson, 2010: ix). Inclusion of both versions of the same dictionary meant that a rough inclusion date could be estimated for neologisms that appeared in *ODO* (as at 31 August 2014) but not *ODE* (published in 2010).

A key criterion in the choice of the final array of expert-produced dictionaries was their perceived relationship with corpora, since corpora are now standard tools for lexicographers, but collaborative *Wiktionary*[20] does not use them. Thus, as explained in 1.1, the dictionaries used in this study would be broken down into the following categories:

- 'corpus-based' (created using mainly corpus data)
- 'corpus-informed' (created using mainly Reading Programmes and citations)
- 'collaborative' (created through collaboration with and between users)

The dictionaries selected were:

- Corpus-based dictionaries:
    - *Oxford Dictionaries* online (*ODO*) (2014)
    - *Oxford Dictionary of English* (*ODE*) (Printed book (2010))
- Corpus-informed dictionaries:
    - *Oxford English Dictionary* (*OED*) (online) (2014)
    - *Merriam-Webster* (online) (2014)
- Collaborative (corpus-free) dictionary:
    - *Wiktionary* (online) (2014)

This choice of dictionaries allowed both the expert-produced versus *Wiktionary* dynamic, and the 'corpus-relationship' to be explored at the same time, since in modern-day dictionary-making, only expert-produced dictionaries are in any way

[20] https://en.*Wiktionary*.org/wiki/*Wiktionary*:Main_Page

connected with corpora, and thus *Wiktionary* stands apart from both sets of criteria. For expert-produced dictionaries, the decision was taken to use almost exclusively Oxford University Press (OUP) publications because these offered the greatest chance of being able to date entries' first appearance. Aside from the online version of *OED*, which includes a 'Publication History' box for many entries (containing not only the initial publication date, but also the dates of any subsequent amendments to the entry) this kind of information is not usually available for expert-produced dictionaries. They are regularly updated (Weiner 2009: 401), but no indication is given in the dictionary itself of what additions or changes were made when. Given that date is such a key element of this study (see 3.4.4 and 3.7.1) this was felt to outweigh any potential bias caused by the fact that all of the British English expert-produced dictionaries in the study were published by OUP. Thus all but one of the dictionaries in the study (*Merriam-Webster*) provided some form of date information against which the inclusion of neologisms could be assessed. However it should be noted that although the 'Publication History' box is present on most entries in the online version of the *OED*, the *OED* dating function is subject to errors, as I discovered on examining the 'Publication History' for the neologism 'greenwashing', shown in Figure 3.1.

Figure 3.1: *OED* entry for 'greenwashing', with 'Publication History' marked

'Greenwashing', was present in *OED* online in August 2014, yet the 'History' claims it entered the online edition in March 2016, as shown in Figure 3.2.

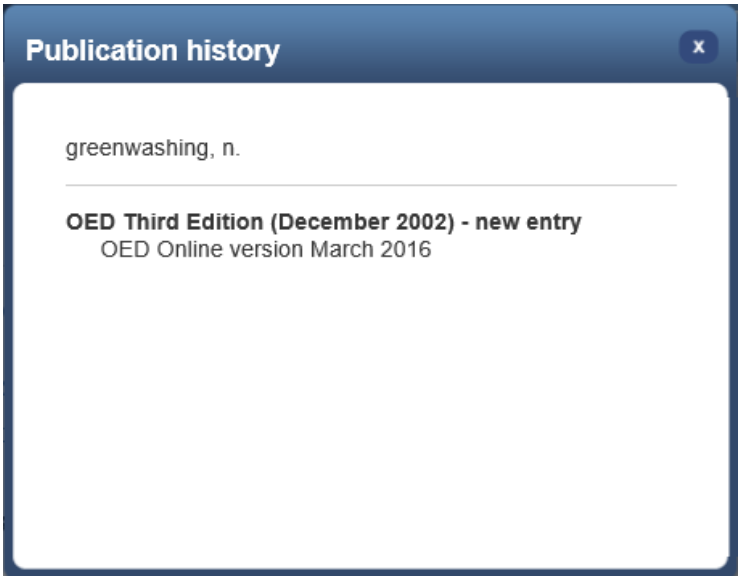


Figure 3.2: *OED* 'Publication History' for 'greenwashing'

Sometimes the *OED* publication history box is completely empty, possibly because this is a new feature still in process of being populated by *OED* staff. This will be discussed further in Chapter 5.

3.4.1 Corpus-Based, Corpus-Informed and Collaborative Dictionaries

In this section I provide an overview of the dictionaries used in this study, in particular how they are created and kept up-to-date, and their relationship (or otherwise) to corpus data.

Although it is recognised that no OUP dictionary is based exclusively on corpora (OUP Principal Language Engineer, Dr Pete Whitelock, Personal Communication 2016), for the purposes of this study the *ODE* and *ODO* are considered to be 'corpus-based', because they are publicised as having been compiled largely from the *Oxford English Corpus* (*OEC*)[21], a 2.5-billion-word corpus created and run by OUP (Oxford University Press 2016d; Oxford University Press 2016e). OUP publicity material states that *OEC* is 'a carefully balanced collection of English for the period of 2000-2006' (Oxford University Press 2016f). In practice, there is data in the corpus dating up until at least

[21] https://en.oxforddictionaries.com/explore/oxford-english-corpus

2011 (see for example the concordance in Figure 3.3), and Whitelock states that most of the data collection was actually done in 2005/6 (Personal Communication 2016).

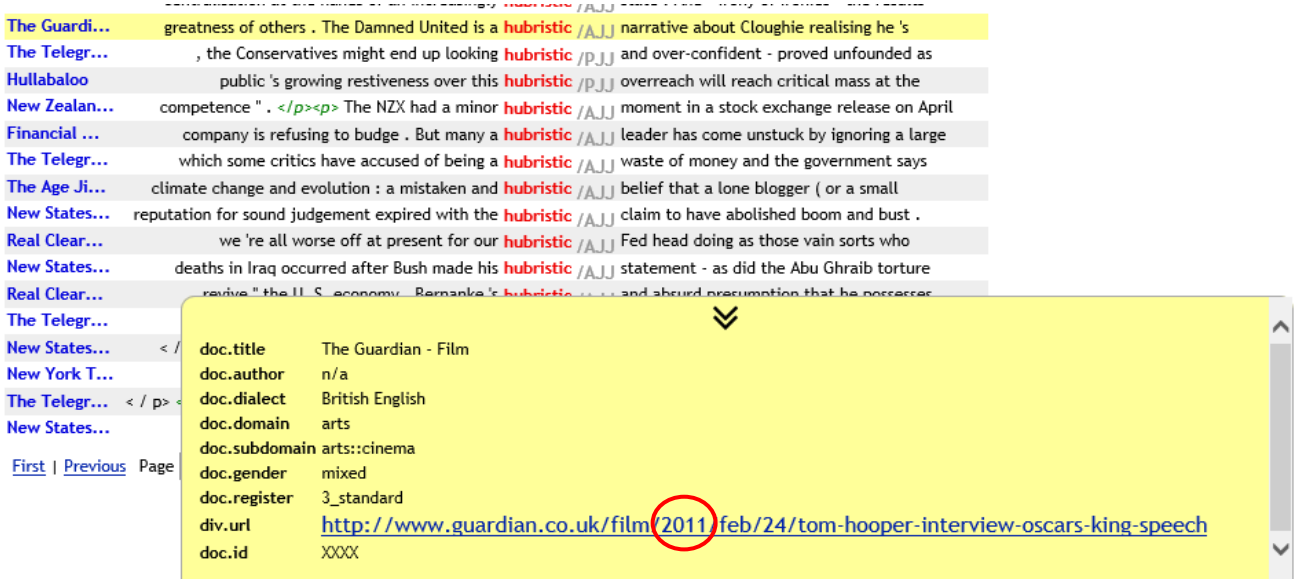


Figure 3.3: Concordance information for 'hubristic' from the *Oxford English Corpus*, drawn from a 2011 *Guardian* article

Compiling a dictionary from a corpus includes developing from corpus data the dictionary components discussed in 3.4.3, such as headwords, definitions, word senses, register and collocation information, phrases and examples (Kilgarriff 2013: 77). On the other hand dictionaries 'informed by' corpora contain some corpus information (for example information about register and collocations) but also information from citations and reading programmes, following the older method of dictionary production, used before the development of corpora (see for example Mugglestone 2011: 51).

Information for inclusion in *OED* is still collected through the dictionary's 'Reading Programme', although editors also have access to the *OEC* for evidence of new words which have become sufficiently well established to be included in the dictionary (see below for dictionary inclusion criteria) (Oxford University Press 2016c; Whitelock, Personal Communication 2016; Oxford University Press 2016h). While initially the Reading Programme was largely confined to literary texts, it now covers sources from all genres of text. A team of readers examines these sources and provides *OED* editors and lexicographers with citations and quotations demonstrating how each word is

used; from here, a definition of that word is devised. Originally, quotations collected by readers were supplied and kept on slips of paper. However they are now collected and held in a comprehensive database, to which all Oxford lexicographers can refer. This database also now includes contextual information which allows lexicographers to analyse words for issues such as subject or date (Oxford University Press 2016c). Words are never removed from the *OED*, and meanings appear chronologically, depending on when each one first appeared (Oxford University Press 2016d).

ODE and *ODO*[22] are dictionaries of current English language usage, with priority given to the most widely used meanings of a word, rather than the oldest (Oxford University Press 2016d). The *Introduction* to the *ODE* suggests that information from the Oxford Reading Programme is used in the dictionary, and the webpage explaining how words enter the *ODO* mentions that the Reading Programme is one of the resources used. However the publicity material (including the book's own back cover, and especially the *ODO* electronic edition) focuses much more heavily on the *OEC*, suggesting that this is the major source of data for the most recent versions of this dictionary (Stevenson, 2010: ix-x, Back Cover; Oxford University Press 2016d). The latest print edition of *ODE* (the 'New' having been dropped from the title for the 2003 second edition) was published in 2010, and an iPhone App version was first released on 26 June 2015 (the App was not used in this study as it was published after the end-date for dictionary inclusion of 31 August 2014.)

A large proportion of the publications used in the *OEC* appear to be US-based or using specifically American English, for example, the *New York Times*, *CNN*, the *Washington Post* and the *Boston Globe.* The only UK national newspapers I have been able to find in the corpus are *The Guardian* and *The Telegraph*. None of the other newspapers used in the current study seem to be included (the *Independent*, the *Mail* or the *Express*). There also appears to be a lot of information missing; for example neologisms found during the course of this study are not necessarily present (or present in the same numbers) in the *OEC* (based on research findings derived from the *Oxford English Corpus*, Oxford University Press) (see 5.3.2 in findings). Unfortunately,

[22] http://www.oxforddictionaries.com/

no-one at OUP has been in a position to explain these limiting factors, despite being extremely helpful in other areas.

As mentioned above, the *OEC* has now been closed down; it was followed by the 7-billion-word *Oxford New Words Corpus*[23], which began in early 2012 and focusses specifically on the appearance of new words in the English Language. It will be interesting in the future to see how this newer corpus develops, and to compare it with both *OEC* and my own database of neologism usage. At present, however, the corpus is for internal use only (Personal Communication, Whitelock 2016).

The publicity material for the American English *Merriam-Webster* dictionary[24] suggests that like *OED* it is not 'based on', but is 'informed by' a corpus (Mitchell 2008: 33-4; Merriam-Webster 2015c). Although *Merriam-Webster* is a dictionary of American English, it was chosen for this study of British English neologisms to see whether there is any delay in the take up of new words in a dictionary covering a different variety of English, and indeed whether there are any gaps, where new British English words simply do not gain a foothold in US English, presumably because the frame of reference needed to understand them does not exist. For example, it is unlikely that the neologism 'dilscoop' (one of the words which could have been chosen for use in this study, and which means 'a certain cricket move ("batting stroke")') would ever gain a footing in American English, due to the lack of interest in cricket as a sport (*NeoCrawler* list, Ludwig-Maximilians Universität n.d.).

Merriam-Webster has been a leading US dictionary for the past 150 years (Merriam-Webster 2015a). Like the Oxford dictionaries, there are a range of *Merriam-Webster* publications; the one used online is the *Merriam-Webster's Collegiate® Dictionary, Eleventh Edition* (Merriam Webster 2015b). According to its website, the *Merriam-Webster* editors operate a system similar to that of *OED*, called 'reading and marking'. This programme collects:

> New usages of existing words, variant spellings, and inflected forms – in short, anything that might help in deciding if a word belongs in the

[23] http://www.oxforddictionaries.com/words/oxford-new-words-corpus
[24] http://www.*Merriam-Webster*.com/

> dictionary, understanding what it means, and determining typical usage. Any word of interest is marked, along with surrounding context that offers insight into its form and use (Merriam-Webster 2015c).

These marked-up references are termed 'citations', and although the collection system dates back to the 1880s, they are now uploaded onto a computer database, to allow for easier access (Ibid). This database contains more than 70 million words and 15 million examples of words. Words are removed only during major revisions (once every 10 years) and those 'tend to be real antiques' (Mitchell 2008: 34). The 'Help' website claims that this database of citations is a corpus (Merriam-Webster 2015c). However there is no indication that any of the key corpus tasks discussed in 3.1 are being carried out, although it may simply be that this information has not been included in the publicity material. Certainly true corpora are 'planned' and 'designed for some linguistic purpose' (Hunston 2002: 2). Merely digitising a collection of citations does not a corpus make; instead, this would be more an archive for the storage of linguistic material, a distinction further made by Hunston (Ibid). Thus in a reversal of the situation with *Oxford Dictionaries* online and the *Oxford Dictionary of English*, the main focus of *Merriam-Webster*'s publicity material's explanation of how the dictionary is developed, is the use of citations (Merriam-Webster 2015c). The fact that the collection system itself has not changed in 150 years, but has simply been digitised, also suggests that the electronic collection of citations is more of a database than a 'corpus'. For this reason, for the purposes of this study the *Merriam-Webster* dictionary, like the *Oxford Dictionary of English*, is deemed to be 'informed by' corpora, but not 'based upon' them.

While it may have no relationship with corpora, collaborative dictionary *Wiktionary* has similarities to expert-produced dictionaries, for example in terms of its language, presentation, functionality and procedures, that make it particularly well-suited to a comparison with expert-produced dictionaries, whatever their 'corpus status'. Of all the collaborative dictionaries available, *Wiktionary* is the one with the most well-defined rules regarding what type of words may be included, how entries are required to look, and how users must go about making changes to the site (Meyer and Gurevych 2012: 273-4). *Wiktionary* also provides detailed help in creating 'proper'

dictionary entries (in the sense that they resemble 'normal' dictionary entries, although they do not actually fit standardised models of a dictionary entry (Wiktionary 2016c; Atkins and Rundell 2008: 200-255, 385-46)). This help, however, is buried deep within the *Wiktionary* website, and from personal experience I believe it quite possible for new contributors to find the 'sandbox'[25] and start practising working on pages, before finding those guidance pages which would enable them to ensure that elements like part of speech, examples, related words and pronunciations are presented relatively consistently (if not constantly). However these elements do not necessarily meet the industry accepted standards for dictionary components discussed in 3.4.3.

One of the key ways in which *Wiktionary* deviates from the model of expert-produced online dictionaries lies in the transparency it provides the user. Not only does each word carry the date it was first included in the dictionary, each entry has a 'Revision History' page which includes every 'save event' ever made on that page (Meyer and Gurevych 2012: 274). This means that every addition, change or deletion to that entry is recorded and can be accessed and viewed by any user. This is also the reason why *Wiktionary* is updated hundreds if not thousands of times a day; every time a contributor presses 'save' on an entry the entire site updates.

There is no committee or body in authority over these changes and updates. As a collaborative dictionary, *Wiktionary* operates on a consensus basis, each contributor being considered an 'editor' and having equal rights and responsibilities to uphold the aims and policies of the website (Wiktionary 2016d; Wiktionary 2016e). That said, rules are in place establishing criteria for inclusion of new words in the dictionary (see 3.4.2 for a full discussion of these criteria). In addition, guidance is provided on how dictionary entries for words meeting these criteria should look, with information both on what is required, and what is preferred of a dictionary entry page. For example the headings that should be used, the information that should be contained within each entry and the way it should be presented (Wiktionary 2016c). The tone is very much one of advising rather than dictating rules, and indeed the site invites deviation from

[25] See https://en.wiktionary.org/wiki/Wiktionary:Sandbox

the suggested formats, but warns that one must be prepared to fight for those changes if other editors disagree[26].

Thus users or researchers can track exactly how an entry developed within *Wiktionary*. The closest any of the expert-produced dictionaries come to this is *OED*'s previously described 'Publication History', as well as notes on when the entry was updated. There is no information, however on what changes were actually made; in *Wiktionary*, the old page is available for immediate review in perpetuity.

In addition to the 'Revision History', *Wiktionary* also has two main Discussion Forums within *Wiktionary*. One is the Tea Room, the central discussion point, where users are directed to open up discussion threads on any entries on which they wish to share thoughts or seek advice. The other, Talk pages, are ostensibly attached to individual dictionary entries (for example 'upskill') but can actually only be accessed by searching the 'Discussions' or 'content' sections of the Community Portal archive (accessible by first searching the Tea Room archives[27]). This is another instance in which information is buried too deeply on the site to be of use unless one is purposefully exploring the *Wiktionary* site. The Tea Room, meanwhile, can occasionally become the site of non-linguistic disagreements, often beginning with a language question (possibly from a non-native speaker) which degenerates into a political or religious argument (see for example 'enemy combatant' from May 2012[28]). It is difficult to know, however how many of these are genuine disagreements and how many are the result of internet 'trolling', (posting 'inflammatory material' in order to try and cause arguments for self-gratification or to try and intentionally disrupt online communities[29]) since they often involve non-registered users of *Wiktionary*, hence the only identification information available is their IP address.

Reviewing lists of Tea Room discussions, however, it appears that such problems are in fact limited, and setting them aside, in my opinion, it is the rules and regulations employed by *Wiktionary* which lead to the perception of it as the most reliable,

[26] https://en.Wiktionary.org/wiki/Wiktionary:Entry_layout
[27] See https://en.Wiktionary.org/wiki/Wiktionary:Tea_room
[28] See https://en.wiktionary.org/wiki/Wiktionary:Tea_room/2012/May#enemy_combatant
[29] https://en.wiktionary.org/w/index.php?title=troll&oldid=28430392

standardised (in terms of its own presentation, permissible language, functionality and procedures) and influential of the collaborative dictionaries. This has led to a number of academic studies on *Wiktionary* as an alternative to 'standard' online dictionaries: see 2.3.4 for example Meyer and Gurevych 2010 and 2012, Penta 2011, Abel and Meyer 2013 and Lew 2012. These factors also mean that a comparison between expert-produced dictionaries and *Wiktionary* would be as close as is currently possible to comparing 'like-with-like' in terms of dictionary entries for new words entering the lexicon. They might also help to facilitate reproduction of the study at a later stage.

3.4.2 Dictionary Inclusion Criteria

In this section I outline the rules governing how new words gain acceptance into each of the dictionaries used in this study.

Although these rules are significantly more relaxed for *Wiktionary*[30] than for the *Oxford English Dictionary* (*OED*) or other 'traditional' dictionaries[31], they are clear and well-thought out, providing users with everything they need to create and refine new entries. Unlike expert-produced dictionaries, in order to gain entry into *Wiktionary*, a word does not have to have been in existence for an extended period of time, or have built up a large collection of citations, but it does have to have proven uses over at least a year. Making words up simply to include them in the dictionary is not allowed in *Wiktionary*, although words which have not yet 'gained acceptance' (which probably means that they have not met the criteria for inclusion), can be included in a special list of 'protologisms' or prototype words. These are kept separate from the dictionary proper[32]. (*Wiktionary* holds an Appendix of Protologisms[33]; none of the words categorised in this study as 'not yet appearing in *Wiktionary*' appear on it.) Any entry which does not meet the site's criteria for new entries can be removed by Administrators (Wiktionary 2016a; Wiktionary 2016b).

[30] https://en.*Wiktionary*.org/wiki/*Wiktionary*:Criteria_for_inclusion
[31] See for example http://www.oxforddictionaries.com/words/how-do-new-words-enter-oxford-dictionaries
[32] See https://en.wiktionary.org/wiki/Wiktionary:Protologisms
[33] See https://en.wiktionary.org/wiki/Appendix:List_of_protologisms

The main rule guiding the entrance of a word into *Wiktionary* is that it 'should be included if it's likely that someone would run across it and want to know what it means' (Wiktionary 2016a). New words in *Wiktionary* need to be 'attested', that is, evidence provided of their widespread use, in 'permanently recorded media, conveying meaning, in at least three independent instances spanning at least a year' (Ibid). The term 'independent' means that the new word must appear in 'different sentences by different people'; it is not enough for one instance of the word to be quoted elsewhere, or for the word to simply be reused by the same person (Ibid). These rules are far more relaxed than those for any of the Oxford dictionaries, or for *Merriam-Webster*. The attestation processes for each of these requires many more citations of the new word, and most require that it have been in use for much longer than a single year, as Mitchell (2008) points out.

Merriam-Webster (*MW*) requires that a word has 'enough citations to show that it is widely used' (Merriam-Webster 2015c). Of course the number of citations is not the only criteria; the number of publications and the length of time they have been used is also crucial: 'a word must be used in a substantial number of citations that come from a wide range of publications over a considerable period of time' (Ibid). This is quite vague; for example, what counts as 'a considerable period of time'? It appears this vague language is due to the fact that the amount of time is very much dependent upon the word itself. According to Mitchell, 'google' entered *Merriam-Webster* as a verb in just five years (the shortest time taken from first citation to inclusion) whereas 'malware' had been monitored since 1990 and at his time of writing was still not included in the dictionary (2008: 33). ('Malware' does now appear in *Merriam-Webster* online, although how long it has been included is unknown, since the dictionary does not provide inclusion dates for entries.)

The *Oxford English Dictionary* (*OED*) is, as discussed above, largely based upon a citational system of data collection. Mitchell cites Jesse Sheidlower, Editor at Large of the *OED* (in 2008) as saying that 'the *OED* is less conservative than *Merriam-Webster* in how quickly it adopts new words' using 'blog' as an example, saying it entered *OED* just four years after its appearance was first noted (2008: 33). The *OED*'s website information on inclusion criteria is less clear. On *OED*'s 'Frequently Asked Questions'

(FAQ) page, in answer to the question 'How does a word qualify for inclusion in the *OED*?' it states that:

> The *OED* requires several independent examples of the word being used, and also evidence that the word has been in use for a reasonable amount of time. The exact time-span and number of examples may vary: for instance, one word may be included on the evidence of only a few examples, spread out over a long period of time, while another may gather momentum very quickly, resulting in a wide range of evidence in a shorter space of time' (Oxford University Press 2016h).

It adds that words should have reached the point where they are 'unselfconsciously used with the expectation of being understood' indicating that no explanation of the meaning of the word is provided (Ibid). Yet just two paragraphs above this on the FAQ page, in answer to the question 'I've made up a word. Please add it to the *OED*', it states that in order for a word to be included in the dictionary, there must be an 'accumulation of a large body of published (preferably printed) citations showing the word in actual use over a period of at least ten years' (Ibid). One of the reasons for such a circumspect approach to inclusion of a new term is likely the fact that once a word enters the *OED*, unlike other dictionaries, it is never removed. This is due to the *OED*'s nature as a historical record of the English language, and not just an indicator of current language usage (Algeo 1993: 283).

However it cannot be ignored that the two answers seem somewhat contradictory, perhaps because in a commercial publishing environment FAQ answers are often written by several different members of a team. There are further differences when we examine the inclusion criteria for *Oxford Dictionaries* online (*ODO*) (and, at the same time the *Oxford Dictionary of English* (*ODE*), since *ODO* is its internet counterpart). The *Oxford Dictionaries* website (which appears to serve all dictionaries except *OED*) states that once there is evidence of a new word being used 'in a variety of different sources … it becomes a candidate for inclusion' in an Oxford dictionary (Oxford University Press 2016h). According to the website, this used to happen over a period of two or three years, yet in the modern digital age, readers expect words

which have become familiar in a very short space of time to also appear in their dictionaries (Ibid).

The *OED*, then, is much more considered in its dictionary inclusion criteria than its cousin the *ODO*, which in turn would appear to be becoming more like *Wiktionary*, in terms of the timings involved in allowing entry of a new word. However *ODO* still requires more evidence of use before allowing inclusion of a new word than does the collaborative work, and of course it is still largely based upon information from the *Oxford English Corpus*.

3.4.3 Standard and Non-Standard Dictionary Components

In this section I outline the components found in the dictionaries used in the current study, which will form one of the central elements of the comparison of degrees of comprehensiveness between expert-produced ('corpus-based' or 'corpus-informed') dictionaries, and *Wiktionary*. The components in the list below are widely recognised as industry standards and are presented by Atkins and Rundell in their 2008 guide to the processes involved in building a dictionary. However although *Wiktionary* may include some of these elements it 'has no fixed structure for its entries' (Meyer and Gurevych 2012: 268). Thus not only does it include additional components not found in expert-produced dictionaries (see Table 3.1) it also adopts a more flexible approach to the 'normal' elements shown here. The impact of this on its dictionary entries (known as 'articles' (Ibid)) will be discussed in Chapter 5 in light of the findings of this comparison.

<u>Industry-standard components of dictionary entries</u>

<u>Headword:</u> (or 'lemma'): An indicator of how a word is written, introducing the rest of the dictionary entry (Atkins and Rundell 2008: 204). According Meyer and Gurevych, in *Wiktionary*, the headword of an entry (known simply as the 'title' of the article) generally appears below an indication of the language of the entry. These 'titles' are case-sensitive, meaning there are different 'articles' for 'café' and for 'Café' (Meyer

and Gurevych 2012: 269, 268). Meyer and Gurevych also discuss 'lexemes', which comprise the headword plus its part of speech (Ibid: 278).

Lexical unit: A subdivision of the meaning of a headword, generally known as a 'sense'. Polysemous entries carry multiple lexical units which can belong to single or multiple word classes (see below). A 'monosemous' headword is both 'lemma' and 'lexical unit' in one, and belongs to one word class (Atkins and Rundell 2008: 204-5). Meyer and Gurevych do not define word senses. They do mention that each sense 'is described by a short 'gloss' (see below) that is sometimes accompanied by references or examples of usage (2012: 282).

Menu: A list of the lexical units (senses) in an entry, although this is largely a feature of learner dictionaries (Atkins and Rundell 2008: 204). Although Meyer and Gurevych make no mention of menus, it is my experience that *Wiktionary* features not a list of 'lexical units' but in many cases a lengthy and detailed navigation pane filled with hyperlinks to help the user move round the 'article' (see Table 3.1).

Definition: An explanation of the meaning of the headword. Definitions are a somewhat contentious issue in lexicographical circles, however, since there are a number of different ways in which a word can be defined (Atkins and Rundell 2008: 405-7). Definitions have multiple uses, for decoding (to understand what one has read or heard) and encoding information (to speak or write) (Ibid: 407-10), and some defining styles are more suited to some users than to others, for example learners of the language versus native speakers (Ibid: 411-13).

The traditional model for English dictionaries is the genus-differentiae model, in which a superordinate term positions the headword in the correct semantic category and additional information indicates what sets it apart from the rest of that category (Ibid: 414). Thus in the definition of a surgeon 'a doctor who does operations in a hospital' (*Oxford Advanced Learners Dictionary* 2005), 'doctor' is the genus (a surgeon being a type of doctor) and 'who does operations in a hospital' is the differentiae, since not all doctors are qualified to perform operations (Atkins and Rundell 2008: 436). This is

generally considered an effective defining strategy, and most dictionaries will use it to some degree (Ibid: 415).

An alternative form of defining is by synonym, where a word meaning the same thing as the headword is used to define it (Ibid: 420-1). However this does not actually explain the meaning of the word. This strategy is only really considered successful when the headword (the 'definiendum') and the synonym 'are semantically identical', a situation which is rare outside of technical contexts, for example 'nanometre' and 'millimicron' (Ibid: 421; Svensén 2009: 215). In most cases, it is considered a less than satisfactory method of defining a word (Rundell and Atkins 2008: 421). Another successful defining strategy is the 'full sentence definition', in which the definiendum is embedded in a complete sentence (Ibid: 441-2). This was the style originally adopted by the COBUILD team as a means of better 'representing what corpus evidence suggested about meaning'. Thus definitions like the following appeared: 'Something that is **immaterial** is not important or not relevant to what you are talking about', 'immaterial' being the definiendum (Moon 2009: 448). This less formal defining style was designed to answer the question 'what does this word mean' in a 'more natural-sounding way', although it did have a tendency to result in long-winded definitions, which were problematic in printed dictionaries (Atkins and Rundell 2008: 208; Svensén 2009: 239). The definiendum-embedded full sentence definition used by COBUILD was effective for nouns, but less so for verbs; here, an 'if' strategy was often used, for example 'defeat: if you defeat someone, you win a victory over them' (Svensén 2009: 237, 239). Full sentence definitions are often used in learner dictionaries, as are the related 'when definitions' which present a shorter definition by using a single clause and no main verb (Atkins and Rundell 2008: 441, 443-4). One example, from the *Longman Essential Activator* (1997) is the definition for 'peace: when there is no war' (Ibid: 443). This style is said to bear similarities to folk-defining techniques, or the ways in which parents and teachers explain the meaning of words to children (Ibid: 444). This approach was not entirely successful when transferred to a written context, however, since the use of 'when' suggested to readers that a main clause was about to follow, leading to confusion when it failed to materialise (Ibid).

Meyer and Gurevych do not mention definitions as such, although they do refer to 'glosses', which they say explain word senses. Based on the examples they use (such as the then third sense for 'boat' (now fourth)) they are discussing the same element (2012: 270).

Pronunciation: Symbols indicating how a word should be pronounced, usually through use of the International Phonetic Alphabet (IPA)[34] (Atkins and Rundell 2008: 206). Meyer and Gurveych claim that *Wiktionary* 'articles' use either IPA or SAMPA (Speech Assessment Methods Phonetic Alphabet)[35] (2012: 269). However although non-IPA pronunciation is found, *Wiktionary*'s guidance pages make no mention of SAMPA[36] and indeed it is not clear if the non-IPA pronunciation symbols do fall under this heading. Non-IPA symbols are also found in the *Merriam-Webster* dictionary[37].

Etymology: The origin of a word and how it has developed through time (common in monolingual dictionaries but rare in learner dictionaries (Atkins and Rundell 2008: 208)). In practice, this often includes word formation processes, as shown in the *Oxford English Dictionary* entry for 'conurbation'[38], although in my analysis I combine etymology with indication of earliest use. Meyer and Gurevych offer a similar explanation to that of Atkins and Rundell (2008) (2012: 269).

Spelling variant: An 'alternative spelling or slight variation in the form of this word' (Atkins and Rundell 2008: 206). This is not something Meyer and Gurevych mention, and indeed the *Wiktionary* guidelines comment only on the need to use correct spelling, since slight differences can indicate a completely different word, using the example 'breath' versus 'breathe'[39].

Word class: The notation showing the word class or part of speech of each lexical unit, for example 'noun', 'verb' 'adjective', often abbreviated to 'n', 'v', 'adj' in print

[34] See for example: https://www.internationalphoneticassociation.org/content/ipa-chart
[35] See for example: http://www.phon.ucl.ac.uk/home/sampa/
[36] See for example: https://en.*Wiktionary*.org/wiki/*Wiktionary*:Pronunciation
[37] https://www.merriam-webster.com/
[38] http://www.oed.com/view/Entry/40647?redirectedFrom=conurbation#eid
[39] See https://en.*Wiktionary*.org/wiki/*Wiktionary*:Quotations

dictionaries, where space is at a premium (Atkins and Rundell 2008: 219). Similar information is provided in *Wiktionary* (Meyer and Gurevych (2012: 270, 280).

Grammar label: Grammatical information about the headword or lexical unit, such as whether a verb is transitive or intransitive (Atkins and Rundell 2008: 221) or whether a noun is only used in its plural form. Meyer and Gurevych do not include 'grammar' in the labels they discuss, however they do make mention of this kind of information (2012: 270).

Register, style and attitude label: Indicators of the type of word, for example 'informal' (register), 'funny' (style) and 'pejorative' (attitude) (Atkins and Rundell 2008: 228-30). Register labelling includes subsets for 'offensive terms' and 'slang and jargon', however Atkins and Rundell note that in some dictionaries the latter are treated relating to register (Ibid: 228). This component, then, is less standardised than the others discussed here. *Wiktionary* uses the same labelling system, although the inclusion of slang and jargon makes up around 40% of register labels (Meyer and Gurevych 2012: 289). As a result of this and the kinds of inconsistencies mentioned above, for the purposes of this study I combine these three label types into one component, known as 'register/style/attitude label'.

Domain label: A marker of the field to which the headword applies, or the context in which it is generally used (Atkins and Rundell 2008: 227; Meyer and Gurevych 2012: 287-8).

Region label: An indicator of where the word is most commonly used (Atkins and Rundell 2008: 227). This is not mentioned by Meyer and Gurevych (2012), although it is listed in the site's guidance information as one of the components of 'context labels'[40].

Example: Text used to elucidate meaning, illustrate contextual features, or attest to the presence of the headword in the language at large (Atkins and Rundell 2008: 453-4). Examples elucidating meaning can clarify differences between senses of polysemous words (Ibid: 454). They can also complement definitions by illustrating

[40] See https://en.*Wiktionary*.org/wiki/*Wiktionary*:Context_labels

usage and providing information on issues such as collocations, syntax and register (Ibid). This can be particularly useful in learner dictionaries, where in order to fully understand how a word is used the user needs to see the different elements in the dictionary entry come together in practice (Ibid).

Examples used as part of the attestation process (mainly found in historical (here, 'corpus-informed') dictionaries like the *OED*) generally come from authentic sources and as such often take the form of quotations (Ibid: 453). Indeed for the purposes of standardised dictionary components, the term 'example' can be assumed to also encompass both 'citations' and 'references'. These quotations are gathered from 'large citation banks', collections of the citations that have contributed to the creation and expansion of the dictionary itself (Ibid: 455). They are attributed, and hence information about source and date can be included in the quotation that appears in the dictionary (Ibid). Other ('corpus-based') dictionaries do not usually include this kind of source information, and examples may come from 'a range of sources (authentic texts, the lexicographer's imagination, or some combination of the two)' (Ibid). The debate over whether examples should be created or extracted from 'authentic' text dates back to the beginning of the COBUILD era, and continues in varying forms to this day (Ibid: 456, 458). Fox raised the issue in 1987 of needing to 'reconcile the requirements of authenticity and typicality' (138). 'Authentic' examples have 'actually occurred in the language' while 'typical' ones show how people regularly use the language (Ibid: 143, 139). A word use can be authentic yet appear only once in real-world language; this does not necessarily make a good example. 'Typical' examples contain uses of words/phrases that repeatedly recur, and so indicate how people typically use the item in their speech and writing (Atkins and Rundell 2008: 459). Yet typical examples may not provide the information that lexicographers believe users need. One way around this is for lexicographers to modify authentic examples, as noted by Laufer in her 2008 study comparing the effectiveness of authentic and modified examples. Another is to use a computerised system to search a corpus and identify what are considered to be 'good' examples of headwords. One such system is GDEX (standing for 'good examples') which was built around the Sketch Engine corpus query program. After inputting the requirements, GDEX presented lexicographers with

a short-list of concordance lines representing potential examples; thus rather than searching through hundreds of concordance lines, they instead had only to select from a shortlist of perhaps 20 (Kilgarriff et al 2008: 1).

In *Wiktionary*, attestational examples (termed 'references' by Meyer and Gurevych 2012: 271) are used, and where there are no suitable quotations, users are encouraged to create example sentences[41]. In practice however it appears that most *Wiktionary* contributors who are unable to include quotations do not actually create examples.

Usage notes: Notes providing additional information on how to use the headword correctly, found most commonly in learner dictionaries. According to Atkins and Rundell, approaches to usage notes vary by dictionary (and thus in my view we must consider these only partially standardised) and they may appear under various different names, including 'functional note', 'synonyms' and 'metaphors' (2008: 233). This naming presumably depends upon the content of the note. The aim of these notes is to 'tell their users what they need to know, even when this will not fit the model of the traditional dictionary entry' (Ibid). Meyer and Gurevych do not mention usage notes. However the *Wiktionary* guidance pages do, indicating that, as in standardised dictionaries, they should show how a word should be used[42].

Cross-reference: An indicator that more information on the headword is available in another entry (Atkins and Rundell 2008: 238). In electronic dictionaries, these take the form of a hyperlink. Cross references are mentioned by Meyer and Gurevych (2012: 268), being similarly used in the dictionary. One example of a cross reference is a hyperlink to a thesaurus. Electronic dictionaries include the functionality of having either built-in resources or access to external ones, such as a thesaurus created by the same publishers. As this is not a feature provided by *Wiktionary*, it is also not mentioned by Meyer and Gurevych.

[41] https://en.wiktionary.org/wiki/Wiktionary:Entry_layout#Example_sentences
[42] See https://en.wiktionary.org/wiki/Wiktionary:Entry_layout#Usage_notes

Run-on: An indicator of a 'derived form' of the headword, for example 'e-tailing' is a run-on of the neologism 'e-tailer'[43] in the *Oxford English Dictionary (OED).* Run-ons are usually found at the end of a dictionary entry. They can be problematic if used too widely, hence they appear in monolingual dictionaries only under strict circumstances, for example where the 'word form is infrequent' (Atkins and Rundell 2008: 236-7, 397). Meyer and Gurveych (2012) make no reference to run-ons. In the context of this study, run-ons are distinguished from instances where the neologism under study is itself a derivative, for example 'promissory note', which in the *OED* is a derivative of 'promissory'[44] (see 'non-standard components' below). Run-ons are indicative of the kind of 'lexical creativity' discussed by Fischer (1998), Renouf (2007) and Moon (2008), although Atkins and Rundell warn against assuming that every word which can be created through the addition of suffixes like '-less' is automatically included in a dictionary as a run-on (2008: 237).

While expert-produced dictionaries are largely constrained by these industry-standard dictionary elements, *Wiktionary* is not. It is able to include a wide variety of additional information, the list of which can be added to at any time (Meyer and Gurevych 2012: 289-90). Table 3.1 provides a list of the kind of non-standard dictionary components which are largely, but not exclusively, found in *Wiktionary*.

[43] http://www.oed.com/view/Entry/251553?redirectedFrom=e-tailer#eid
[44] http://www.oed.com/view/Entry/152449?redirectedFrom=promissory+note#eid28182018

Non-Standard Dictionary Components (mainly used in *Wiktionary*)	
Inclusion date	Indicator of when the word first entered the dictionary (also provided for some words in *OED*)
Revision History	Save-by-save record of every change ever made to an entry
Discussion Forum	Online spaces for discussion of dictionary entries, specifically Talk pages and the Tea Room
Audio File	Sound file added to help with pronunciation (now found in many electronic expert-produced dictionaries)
Translation	Headwords provided in multiple languages (*Wiktionary* only)
Derivative	Marker that the neologism under study derives from another headword, for example in *OED* 'cyberbullying' is a derivative of 'cyber'
Related term	Indicator that a word is linked to the headword, although it is not an actual run-on. In standardised dictionaries this might appear as a Usage Note, however it is treated separately here as *Wiktionary* treats it as a separate element
Synonym	Word which means the same as the headword. Also often included in Usage Notes, but separated here for the same reason as related terms
Contents navigation panel	Panel of hyperlinks to help users move around longer entries in *Wiktionary*

Table 3.1: Non-standard dictionary components

While some of these elements are similar to standardised elements (for example 'related terms') or indeed found in standard dictionaries in different formats ('synonyms' often being found in Usage Notes (Atkins and Rundell 2008: 33)), others are specific to the collaborative form. These include the revision history and discussion forum.

3.4.4 Dictionary Date of Entry Datasets

Objectives 1 and 2 of this study were to compare degrees of comprehensiveness in the entries provided for new words in expert-produced dictionaries with those in collaborative dictionary *Wiktionary,* and to track neologism appearances in UK news media in order to compare usage and behaviour in different newspapers at different stages in the neologic life-cycle. In order to achieve these objectives it was necessary that the neologisms chosen be organised by date. For Objective 1, this would allow for the comparisons of representations of neologisms in expert-produced dictionaries and *Wiktionary*, to take account of any changes occurring over time (for example, whether entries for a new word remain constant after first inclusion, or whether they are expanded over time). It would be even more important for the tracking of neologism

usage in the media, since Objective 2 specifically requires that this usage be examined at different stages in the neologic life-cycle.

As a result of this, it was necessary to establish 'gradations of newness' for the words across the 14 years of the study, which corresponded with stages in the neologic life-cycle (see 1.1). This would allow a word which had been in use for 10 years to be considered separately from one which had been present for only five. Since this was a lexicographical study, comparing *Wiktionary* with expert-produced dictionaries, the dates chosen for these 'gradations of newness' would be based upon when a word first entered *Wiktionary* or an expert-produced dictionary. The resulting datasets would consequently be termed Dictionary Date of Entry Batches or DDEBs.

Kerremans had established as 'new' those words which 'did not occur before 2006' (2015: 81), meaning that she was studying neologisms over the most recent six-year period (her research concluding in 2011 (Ibid: 115)). Whilst I found the process she used to establish when new words entered use to be flawed, in particular her choice of dictionary to judge the presence of neologisms in the lexicon (see 2.4), my subjective view was that this six-year cut-off point was appropriate. Words more than six years old have passed beyond the point where I, as a consumer of language, consider them to be 'brand new'. In addition, Fischer (1998) had also chosen six years as the period of study for her examination of creative neologisms in *The Guardian* and *The Miami Herald*, suggesting that this was a reasonable time frame.

Neologisms included in this original DDEB would therefore have entered the dictionaries used in this study between September 2008 and August 2014, when data collection began (see 4.5.4). However there would be an additional DDEB comprising neologisms which had yet to enter a dictionary. The third and final DDEB would comprise more established neologisms, which had entered dictionaries some years previously. These would show how we might expect the words in the first two batches to develop over time.

The date parameters for this category were based upon the development of interactivity on the Web, the history of *Wiktionary* and changes to the *Oxford English*

Dictionary (*OED*). As mentioned in 1.1, Web 2.0 brought about a new age of internet interactivity, beginning around 2001 (Neuman, Nave and Dolev 2010: 58). *Wiktionary* was launched in 2002 (Meyer and Gurevych 2012: 261), following *Wikipedia* in 2001 (Bryant, Forte and Bruckman 2005: 1). *OED* first appeared online in 2000, and during the same period took on a new, more proactive approach with regard to neologisms (Weiner 2009: 401-2). As a result, I adopted January 2000 as the starting point for the third DDEB and the earliest point on the neologic life-cycle. New words entered the neologic life-cycle on any date between this and 31 August 2014 (when data collection for the study began). However this did not mean that the neologisms used here were in any way considered established at that point, since this was not the intention of the study. For all practical purposes in the conducting of this research project then, the neologic life-cycle would be broken into three datasets, categorised as follows:

- Dictionary Date of Entry Batch 1 (DDEB1):
    - Neologisms not yet appearing in *Wiktionary* and/or any expert-produced dictionary as at 31 August 2014
- Dictionary Date of Entry Batch 2 (DDEB2):
    - Neologisms entering *Wiktionary* and/or an expert-produced dictionary between September 2008 and August 2014
- Dictionary Date of Entry Batch 3 (DDEB3):
    - Neologisms appearing in *Wiktionary* and/or an expert produced dictionary between January 2000 and August 2008

Since this study was concerned with the date of changes in the representation, use and behaviour of neologisms, in most instances DDEB 1+2 were combined, since they effectively covered the same timeframe. However in some instances it was useful to be able to exclude words not yet appearing in a dictionary (for example when conducting comparisons of dictionary entries) or to separate them from those which have recently entered dictionaries, in order to obtain a more detailed picture (for

example during media tracking). At these times, then, DDEB1 would either be excluded altogether, or treated as a separate category from DDEB2.

3.5 Elements of Project: Newspapers

In this section I move specifically into the 'media tracking' portion of the study, presenting the newspapers selected for the project. I discuss the socio-economic groups to which these newspapers belong and the influence of this on my choices. I also outline the requirement for professional journalistic writing, in order to be able to explore standardised usage of these new words, through appearance in publications subject to corporate control and regulation.

Newspapers were chosen as the most appropriate vehicle to represent media usage of neologisms, since 90% of the population read 'newsbrands' (Newsworks 2016) (the new term for news media across print and digital platforms (Greenslade 2012)) every month. Newspapers continue to be one of the key ways in which most people interact with language, and in particular, new language. It has long been believed, as Facchinetti points out in her article on news writing from the 1960s onwards, that 'the language of news is supposed to be first and foremost factual' (2012: 145). If this is in fact the case, it may be that readers accept new words in these articles as being more reliable and better established than they would if seeing them in a less formalised context, such as the blogs in which they were originally identified and tracked by Kerremans (2015) (see 3.6.2).

3.5.1 Socio-Economic Factors Influencing Choice

The choice of newspapers for this study was based upon socio-economic categorisation of readership, to ensure reach across all sectors of the British population, as defined by social class and economic factors. Newspapers were selected based upon the *National Readership Survey*'s (NRS) 'social grade categories' (see Table 3.2).

Social grade	Social status	Occupation
A	Upper middle class	Higher managerial, administrative or professional
B	Middle class	Intermediate managerial, administrative or professional
C1	Lower middle class	Supervisory or clerical, junior managerial, administrative or professional
C2	Skilled working class	Skill manual workers
D	Working class	Semi and unskilled manual workers
E	Those at lowest level of subsistence	State pensioners or widows (no other earner), casual or lowest grade workers

Table 3.2: NRS social grade definitions (businessballs.com 2015)

The newspapers chosen were:

- *The Guardian* (*Guardian* and *Observer*) (Guardian News and Media 2014)
- *Independent* (Independent.co.uk 2014)
- *Mail* (*Daily Mail* and *Mail on Sunday*) (Associated Newspapers 2014)
- *Express* (*Daily Express* and *Express on Sunday*) (Northern and Shell Media Productions 2014)

The Sun (News Group Newspapers 2014) was also originally included in the list, but was later excluded (see 4.2.2 The *Sun vs Google Advanced Search).*

The requirement of this study was that the newspapers chosen be UK-based and cover national news. *The Guardian* and the *Independent* were chosen in part because they were each included as data sources for both Kerremans' (2012) and Renouf's (2013) neologism-based studies, as well as *The Guardian* being used by Fischer in her 1998 study of creative neologisms in newspapers. It thus made sense to build upon existing 'neologic' knowledge of these publications. Data provided by Ipsos Mori (which conducts the *National Readership Survey*) indicates that *The Guardian* and the *Independent* are mainly read by Social Groups A and B, providing me access to the higher echelons of newspaper readership, whilst the *Daily Mail* and *Daily Express* are

read mainly by those in C1 and C2. (It is not clear whether Ipsos Mori includes the *Mail on Sunday* or *The Sunday Express* in these figures) (Duffy and Rowden 2005: 21). The latter two newspapers had not been previously examined in terms of their approach to neologisms, and counter-balanced the two papers which had a known history of neologism use.

The social grouping classifications shown in Table 3.2 would lead us to expect that readers of *The Guardian* and the *Independent* would have a higher level of education than those reading the *Mail* or the *Express,* usually having had to attend university in order to achieve the managerial and professional jobs they enjoy in social groups A and B. We would also expect the former to display higher levels of political and societal engagement, in line with the contents of these papers (although evidence from Ipsos Mori suggests that the latter differential is less pronounced than expected (Duffy and Rowden 2005 11-14)). The differences between these publications date back to major changes in the newspaper industry in the early 20th Century when 'New Journalism' and 'tabloidisation' saw the beginning of a clear distinction between tabloid newspapers (with shorter stories and lots of captions, aimed at the working classes) and non-tabloids, offering hard news aimed at the higher echelons of society (Bös 2012: 101-105).

In each case, only articles which had appeared in both the print and the online editions of the newspaper were included the current study. Online-only articles tend to be marked as such, for example 'for *Mail Online'* appears below the byline (author's name) in articles which have not appeared in the print editions of either the *Daily Mail* or the *Mail on Sunday.* This decision was made to ensure that the articles being studied had been seen by all of the newspapers' readership, rather than one section or the other.

3.5.2 Professional Journalism

One key criterion for inclusion in this study was that the newspaper articles had to be written by professional journalists, since it was the newspapers' use of neologisms that was under investigation. These neologisms had already been studied within a

social media context, during Kerremans' *NeoCrawler* study (2015), which generated the list of neologisms from which the words used in this study were selected (see 4.2 and its subsections).

It was expected that the work of these professional journalists would be governed by their newspaper's policies towards new words (for example whether/when they should be used in inverted commas and if they should be italicised or glossed in the text). However even following the strictest of style sheets there will always be some differences in individual authorial style. I expected that the journalists themselves had achieved a pre-determined level of education (generally degree or equivalent) and that their written English was of a consistently high standard, with the ability to both inform and entertain readers. These requirements are apparent in a job advertisement for a 'senior editor/journalist' for the global publishing company LexisNexis[45], shown in Appendix 1. Journalists applying for this post are expected to be able to write in a 'clear, succinct' style, aimed at a particular audience which has knowledge and experience of the specific topic. They are expected to be able to conduct interviews, to proof-read and correct copy, as well as writing 'interesting stories for a time-poor audience'. An in-depth knowledge and interest in the field are required, as is a university degree, with a number of specific degree topics mentioned[46].

Confining the current study to writing by these kinds of journalists enabled me to track the standardised usage of these new words in publications governed by strict branding and style guidelines. In addition, the selection of newspapers according to the criteria discussed above was designed to allow for replicability and/or reproducibility of the study (as far as is possible given the constraints of a web-based database) by future researchers.

[45] See https://jobs.theguardian.com/job/6350457/senior-editor-journalist/
[46] See https://jobs.theguardian.com/job/6350457/senior-editor-journalist/

3.6 Elements of Project: Web-Based Corpora

Since the advent of the World Wide Web as a source of data for the building of corpora, corpus projects have been split into two distinct categories: those which use the web itself as a corpus to analyse, and those which take specific types of information from the web and build corpora from there. These two types of project are widely known as 'web-as-corpus' (WAC) and 'web-for-corpus' (WFC) respectively.

The distinctions between the two are significant. The former tend to be much larger, simply because they are utilising the entire World Wide Web as their data source. According to Alpert and Hajaj (2008, cited in Fletcher 2013: 1) in 2008 Google claimed to have identified more than 'a trillion (10^{12}) distinct' Universal Resource Locators (URLs or web addresses) and to have found that 'several billion (10^{9}) new web pages appear every day'. Corpora built <u>from</u> the web, meanwhile, tend to be much smaller, simply because they are more targeted, collecting data from a particular type of English, rather than the whole of the World Wide Web. They therefore often tend to be genre-specific, allowing researchers to individually study a wide range of different language types (from historical English to the language of social media) and lexicographers to create dictionaries of English for Specific Purposes, such as Business, Aviation or Medicine (Hundt, Nesselhauf and Biewer 2007: 10). Such genre-specific corpora include the Zen Corpus and the Rostock Newspaper Corpus, both of which are used to provide linguistic analysis of historical newspapers (Fries 2012: 49-90; Bös 2012 107-144) and Renouf and Fischer's corpora of modern-day English journalistic writing (2013 and 1998 respectively).

More recently, we see corpora such as 'Monco'[47], which was launched (in 'beta development phase') in early 2016, and claimed to be a 'near real-time monitor corpus'. 'Monco' contained 1.1 billion words at the point of launch, with an additional eight million added daily (Piotr Pezik, Corpora Digest mailing list entry, corpora@uib.no, 2016).

[47] http://monitorcorpus.com

This focus on 'real-time' stems from a problem also recognised by Kerremans, Stegmayr and Schmid (2012: 62), who note the problematic 'time lag' that can exist 'between data collection and public access'. By the time a corpus is ready for research a word which was new when it was collected can have become obsolete, or conversely have been thoroughly institutionalised. Speedy identification of new words is therefore crucial in order to allow for effective investigation of the earliest phases of the establishment process (Ibid). The main function of 'Monco' was said to be 'to provide the most recent examples of English usage', allowing users to 'find examples of [such] new lexical items' as well as serving 'as a general purpose reference corpus in its own right' (Piotr Pezik, Corpora Digest mailing list entry, corpora@uib.no, 2016).

This new web-based corpus contained all of the newspapers I use in the current study, and I therefore tested it to see if it might be possible to incorporate it into my own research. However it was found to be extremely unstable (likely due to its beta status). The drop-down menus were in Polish and offered no English translation. Other elements were in English, but did not accept changes, for example to the date range for searching the corpus (to fit the date range of my study: 2000-2014). The results of my searches also varied significantly depending on the neologism, for example my search for 'bankster' would only go back as far as 2015 yet I know from my own data that there were entries in 2012 and 2014.

Due to these problems, the Monco corpus was abandoned as a possible addition to the current study, although it may be that later versions will be of more use to future researchers.

A number of other corpora were also considered as possible additions to the data sources for this research project. COCA, the Corpus of Contemporary American English[48] was rejected on the basis that my study is designed to examine British English, not American English. The Corpus of Global Web-Based English (GloWbe)[49] was considered, but although it contained 1.9 billion words from 20 English-speaking countries, it did not appear possible to specifically search newspapers. In fact the only

[48] http://corpus.byu.edu/coca/
[49] http://corpus.byu.edu/glowbe/

domain which could be independently interrogated was Google Blogs; this had already been covered by Kerremans' original study (see 2.4). As I was specifically interested in the behaviour and use of my set of 34 neologisms in four specific UK national newspapers, the generalised search available through GloWbe was not suitable. It was therefore also rejected.

Finally I considered the newspaper corpus used by Renouf and her colleagues in the various papers discussed in 2.2 and its subsections. Renouf et al accessed this corpus via the WebCorp LSE program[50]. However, although several corpora are available to the public in this way (including a corpus of blogs), the newspaper corpus does not appear to be open access. In addition, Renouf and Kehoe state that the corpus contains UK broadsheet newspapers *The Guardian*, the *Independent* and the *Observer* (2013: 168). As I sought to explore neologisms in newspapers across the range of socio-economic groups (including tabloids the *Mail* and the *Express*), this corpus would not have been suitable even if access had been allowed.

3.6.1 Manual versus Automated Data Collection Methods

Due to the size and extent of the material utilised in web-as-corpus (WAC) projects, automated methods are used to extract the data for corpus analysis from the web; it would be next to impossible to take on such a task manually. These automated programs are generally based on commercially available search engines such as Google, Yahoo and Bing, with freely available corpus query tools attached, for example WebCorp[51] and KWiCFinder[52] (Fletcher 2013: 3-4,). Standalone programs also exist, such as Sketch Engine[53].

While web-for-corpus (WFC) approaches might make it more feasible to collect data manually, this has always been impractical unless the corpus is small in size. For small databases such as mine, of a little over four million words, a degree of automation is useful to facilitate the downloading of target webpages. Automated processes are

[50] http://wse1.webcorp.org.uk/
[51] www.webcorp.org.uk
[52] http://kwicfinder.com
[53] https://thesketchengine.co.uk

generally conducted using 'webcrawlers' or 'spiders' to 'harvest' webpages, starting from 'seed URLs' and following each link on each page 'until user-defined criteria are met' (Fletcher 2013: 5). While there are a number of programs available to complete this task, most, according to Fletcher, require 'programming expertise to customize and coordinate the various processes' (Ibid).

Once harvested, the webpages collected through these processes are prepared for use in corpus analysis programs, being converted to plain text (.txt) files, and removing recurring information such as navigation links, along with duplicate or almost duplicated documents and those comprising non-textual material such as images (Fletcher 2013 :5). Grammatical tagging and correction of spelling or typographical errors may also be carried out (Fletcher 2013: 5). From here, data analysis can begin, using corpus analysis tools such as Sketch Engine, AntConC[54] or Wmatrix[55].

3.6.2 Text Selection and Collection in Web-Based Corpus Studies

While web-as-corpus (WAC) projects do not need to select texts for use in their studies, since they are using the entire web as their data source, web-for-corpus (WFC) researchers do. As discussed above, this is often done using standard webcrawling programs that follow links from webpage to webpage, downloading text as they go.

A newly developed automated system, however, brought these two methods together into a single program, the *NeoCrawler* (see 2.4). The methods and findings of this system will be used as an exemplar of a new phase in automated text selection and collection, and it is against this that my own manual methods will be compared.

The *NeoCrawler* comprised two component parts: the *Discoverer* which searched for first instances of neologisms, and the *Observer* which then tracked their development, all within the confines of the Google Blogs environment. In searching for first coinage of a neologism, the *Discoverer* searched the web much as do standard webcrawling programs, downloading pages and extracting 'those grapheme sequences that are not

[54] http://www.laurenceanthony.net/software.html
[55] http://ucrel.lancs.ac.uk/wmatrix/

contained in the *Discoverer's* dictionary', identifying these as potential neologisms (Kerremans 2015: 80-1, 78-92). The system weeded out 'non-words' ('sequences of letters that resemble words but in fact are not' (Ibid: 82)) and assigned values to potential new words that indicated the likelihood of them being true neologisms. Manual filtering was used during the final stages, and potential neologisms were classified according to the word formation processes that led to their creation (Ibid: 82-3). The *Observer*, meanwhile, operated as a kind of cross between a web-for-corpus (WFC) crawler and a web-as-corpus (WAC) search program. It used Google to conduct the search much as would any human searcher, however the search parameters were preprogramed, allowing the *NeoCrawler* to disguise 'itself as a web browser' (Ibid: 84). As the *Observer* was only interested in contemporaneous instances of neologism use, searches were conducted weekly and were limited to the previous seven days (Ibid)). HTML results returned by Google were 'parsed', in this case an automated process for removing unwanted pages and links (Ibid).

It is against this process of tracking the use of neologisms over time that my own new manual methodology is compared in the current project. The basic procedures are the same: using Google to search the web for appearances of a predetermined set of neologisms. However my manual system is designed to allow for more nuanced tracking. Through the use of 'pre-screening' and 'advance exploration' of websites (see 3.7), it is possible to exclude and narrow down search parameters to a much greater degree than can be achieved through automated searching. This leads to more targeted results, with more context than cotext.

3.7 Aims and Summary of New Methodology for the Collection of Context-Rich Genre-Specific Corpus Data

As mentioned in 3.1, one of the central aims of this project was the development of a new methodology that would allow for the creation of much larger context-rich genre-specific corpora than has previously been the case. It was determined through a process of trial and error that this was possible through the development of more

nuanced manual data collection procedures which would enable important contextual information to be gathered for large quantities of corpus texts. This represents one of the major contributions of this project to academic study.

A key element of this new methodology is the way in which unwanted data is excluded before the database/corpus is created (through 'pre-screening' of search results and 'advance exploration' of webpages) rather than during post-processing when 'noise' (data that does not meet the researcher's selection criteria and could therefore skew the results (Fletcher 2013: 5)) is usually removed. The new approach allows not only for the creation of a more context-rich genre-specific database, but also faster, more efficient data collection that facilitates the building of larger corpora than has previously been the case without access to complex algorithms.

The new methodology, which is explained in full in Chapter 4, can be summarised as follows. Google Advanced Search (GAS) (see 4.3.1) was used to search for neologism usage (including spelling variations (see 4.3.2.1)) within the internet domain of each of the four newspapers:

- The *Guardian*
- *Independent*
- *Mail*
- *Express*

Each search generated Search Results Pages (SRPs) containing hyperlinks to each of the pages containing usage of that particular neologism. Below each hyperlink was a short extract from the target page, which in most cases showed the neologism in place. Initial 'pre-screening' of these search results was conducted based on this information, allowing results to be 'pre-excluded'. Results excluded in this way (see 4.5.1) were:

- False positives, or words similar to the neologism but not correct, for example 'iPad' for 'iPadable'

- Instances where the SRP extract did not contain the neologism
- Duplicated webpages, identifiable through repeated blocks of text on the SRPs
- Neologisms appearing somewhere on the page other than in the main article (identifiable through repeated blocks of text in the SRP indicating, for example, a link to another page)
- Archived articles, containing texts which had already been identified in their original form
- Advertisements not collected since they relate to the user's previous browsing history and their location, rather than the search word
- URLs featuring the file extension '.gz.xml' or comprising the file name 'robots.txt'; these were not newspaper articles and so were not relevant to the study.

Supplementary screening was conducted on first examination of each webpage, and any of the following resulted in the article being pre-excluded from the study (see 4.5.1):

- No date, or a publication date outside of the required date range
- Wrong parts of speech for all neologisms in the article
- Neologisms as a topic for the article
- Article attributed to a press agency such as Reuters[56] rather than to an individual journalist
- Publication exclusively to the online version of the newspaper
- Duplication of an archive or 'round-up' article
- Entire article made up of a verbatim transcript of a speech

[56] http://uk.reuters.com/

- Paid-for or sponsored article
- No text: only photographs or videos
- Broken links, for example a '404 page not found' error
- Missing data due to expiry of copyright/licence
- Internal search results.

Each of the remaining hyperlinks on the SRP was clicked, and on entering the target page, the following information was manually collected and entered into a Microsoft Word document:

- URL (Universal Resource Locator or web address)
- Newspaper title
- Publication date
- Number of instances of neologism and if it appeared in the headline
- Article type (for example 'news')

All of the data in Microsoft Word was converted into a table and transferred into Microsoft Excel, where it became the searchable *NTON* (*Neologism Tracking in Online Newspapers*) database. This version of the database was used for all frequency queries. All of the URLs from the database were uploaded into Sketch Engine, so that concordance lines could be produced showing neologisms in use.

The following explorations were conducted using the *NTON* database:

- The tracking of neologisms over the past 14 years, to examine changes in meaning and behaviour
- Comparison of the treatment of neologisms in different dictionaries.

The new manual method of compiling *NTON* was compared with the automated *NeoCrawler* program to explore which system might be most effective for collecting context-rich genre-specific data.

3.7.1 Key Contextual Information – Date

One of the most important pieces of contextual information in this study was the publication date of each article which contained a neologism. In Kerremans' study of the *NeoCrawler*, it is 'cotext' (words appearing with the neologism) more than context which is important, to the extent that she refers to it as 'co(n)text' (2015: 47-54). In addition, Kerremans defines 'context' as 'an umbrella term for linguistic cotext and extralinguistic context' (Ibid: 47). From a reading of her thesis, I take this to mean the additional language surrounding a neologism which helps the reader to decode and understand the new word. As she says: 'context provides linguistic clues that provide the mental lexicon with valuable input to elicit more precise interpretations and increase comprehension' (Ibid: 52). My own use of the word 'context' is slightly different, referring to the additional information contained within a corpus text which helps to position and explain that text, for example date, author or article type.

Only through collection of the publication dates of these articles could the use and behaviour of the new words be tracked over time. This was also one of the elements which was found to most benefit from the choice of a manual methodology over an automated one, since programming any system to deal with all of the different ways in which a date can be presented is likely to be highly complicated.

Dates can be presented as follows:

- Numerically:
    - including a '0' where a day or month is a single digit
    - excluding a '0' where a day or month is a single digit
- Semi-numerically:
    - with ordinal numbers

    - without ordinal numbers

- with the day appearing:
    - before the month
    - after the month
- With the month abbreviated
- With an abbreviation of the year
- With the year *in toto*
- Non-numerically (not found in newspapers)

Or any combination thereof.

For reasons unknown, a number of articles in any given newspaper carry no date at all, and, rather than a date, online versions of newspapers give the number of hours since the story was released on the day of publication. Dates can also appear in different places on a page, and this can make them difficult to locate once the page has been downloaded from the web and has gone through post-processing, turning it into a run of unformatted text. The date might originally appear:

- Above the headline
- Below the headline but above the byline
- Below the byline but above a set of introductory bullet points
- Just above the start of the actual text of the article.

Finally, styles may differ within individual newspapers depending on the section of the paper, for example dates might appear differently in a Sunday edition than a weekday one.

A bespoke webcrawling programme could have been created (or a standard one customised) to search for each of the 12 named months of the year (plus their 12 standard abbreviations) and their numerical counterparts. However the possibility for confusion exists, for example the byline of a journalist with the first name 'June' could confuse the system into believing it had found a partial date, as could reference to an event in the text itself, for example 'last November's fireworks party'. One mechanism said to be built into the Web which could potentially address some of these difficulties is the 'Last Modified' header, which indicates the last time a webpage was saved. However results from testing this have shown that it is present in only a little over 50% of cases, and hence this was not investigated further here (Kehoe 2006: 297-8).

Through trial and error, it was determined that it would be simpler to access the articles manually, when the date was still clearly visible, code them according to date and then download them. Thus, despite Lüdeling, Evert and Baroni's assertion that 'except for very small corpora, the process of downloading web pages to build the corpus (and any post-processing that is applied) must be automated' (2007: 25), I decided that a manual approach was required for text selection for the *NTON* database.

Of course the development of this new methodology was very much a 'trial and error' process, as will be discussed in Chapter 4, where I outline the major successes, difficulties and solutions which led to the creation of the final new methodology.

3.8 Ethical Considerations

Ethical approval was sought for this project in 2013 through Coventry University's online Ethics Procedure, which is overseen by the University Applied Research Committee[57].

As the data used in this study is all in the public domain (available via search engines), and no living participants are involved (viz, there is no direct contact with

[57] See https://ethics.coventry.ac.uk/about/ethics-at-cu.aspx

lexicographers working on any of the dictionaries under study, or with journalists from any of the four newspapers) the project was awarded Low Risk ethical approval.

3.9 Research Questions

In this section I introduce the three Research Questions which this study seeks to answer, along with a short rationale for each one.

As a lifelong dictionary enthusiast and linguaphile, the changes that have taken place in the field of lexicography over recent decades have fascinated me. As dictionaries have moved online, the amount and type of information they can offer has expanded, whilst at the same time new types of dictionary have appeared (see for example Nesi 2009). Chief among these, for me, is the collaborative dictionary. The idea of ordinary people taking charge of the dictionary-making process fascinated me. I was especially keen to find out whether the approach to new words was the same in collaborative dictionaries as in expert-produced ones, which are created using some degree of corpus input. I therefore set out to compare the two, from the perspective of responsiveness to neologisms and levels of detail in new word entries.

> **Research Question 1 –** *What can be learnt from this study about* Wiktionary's *responsiveness to neologisms and the level of detail and quality of definitions in its new word entries, when compared with expert-produced dictionaries?*

I decided to conduct a Media Tracking project in order to draw a complementary picture of the real-world usage of the neologisms studied in the dictionary context above. This was to allow me to explore the use and behaviour of neologisms at different stages in their 'neologic life-cycle'. This life-cycle, unlike that proposed by, for example Renouf (2007 and 2013), covers **only** the period during which a word is considered 'new' under the parameters of this study (see 1.1), and is indicated by new words' presence in any one of the three Dictionary Date of Entry datasets (see 3.4.4). I believed that there would be a correlation between the stage of 'newness' of

a word, and its presence in the British news media, demonstrating media attitudes towards new words, and their gradual process of acceptance.

> **Research Question 2 –** *What can be discovered about the 'neologic life-cycle' of selected neologisms in UK national newspapers between 2000 and 2014?*

This kind of media tracking would ordinarily have been conducted using a standard webcrawler; such programs search the web, following a trail of URLs and downloading pages as they go (Fletcher 2013: 5). In this case, it would have searched the newspaper domains seeking articles containing the neologisms in question, much as was done by the *NeoCrawler* in tracing neologism usage within the Google Blogs environment. However systems like the *NeoCrawler* provide little or no facility for collecting contextual information from the sites they visit, and it was this kind of information that would be required to draw useful conclusions here. I therefore decided to devise a new manual methodology which could incorporate this additional functionality, and which would hopefully prove useful to future researchers engaged in genre-specific studies where context is key.

> **Research Question 3 –** *In the context of data collection for context-rich, genre-specific web-based corpora, is the proposed new manual methodology more or less appropriate and effective in tracking neologism use and behaviour than the automated methods of the kind used by the* NeoCrawler*?*

3.10 Conclusion

In this chapter I outlined the methodological framework within which the current research project is conducted, and I presented the key elements underpinning that research project, specifically dictionaries, newspapers and web-based corpus data.

In discussing dictionaries, I considered collaborative and expert-produced publications, as well as the relationship between the latter and corpora. I outlined the inclusion criteria which govern acceptance of neologisms into each of the dictionaries used in this study, and I presented the standard and non-standard components that

make up entries in each of the different dictionaries types used here. Finally I established the Dictionary Date of Entry Datasets (DDEBs) into which neologisms in the study are organised.

In my discussion of newspapers, I explored the socio-economic factors which influenced which four I chose to include in my study, and I stressed the importance of professional journalistic writing in order to build a picture of newspapers' attitudes towards neologisms. In discussing web-based corpora, I considered the two different approaches to web-based data collection (web-as-corpus and web-for-corpus) and I presented a number of existing corpora which might have proved useful to this study, but were ultimately rejected for a variety of reasons. I examined current manual versus automated data collection methods, and I introduced the *NeoCrawler*, the automated system against which my own manual methodology was compared.

Chapter 4 Methods and Methodology

Part 2 – Data Collection and Analysis

4.1 Introduction

In this chapter I explain how data was collected and analysed in order to address the three Research Questions established in Section 3.9 of the previous chapter. I discuss the major issues encountered during the course of this project, and present a detailed account of the creation and piloting of the new methodology devised here to track neologism behaviour and usage in UK national newspapers. This was done with a view to achieving the key objectives of the study:

1. To compare degrees of comprehensiveness in the entries provided for new words in expert-produced dictionaries with those in collaborative dictionary *Wiktionary*

2. To track neologism appearances in UK news media in order to compare usage and behaviour in different newspapers at different stages in the neologic life-cycle

3. To consider whether neologism use and behaviour in the media can be best explored through the use of new manual or existing automated corpus data collection techniques

For Objective 1, new words' entries from five different dictionaries were analysed, comparing the number and quality of dictionary components (standardised and non-standardised). This was to determine whether *Wiktionary* is indeed more responsive to neologisms than expert-produced dictionaries, and whether its entries are more detailed. Research Question 1 had been established to explore this:

> **Research Question 1 –** *What can be learnt from this study about* Wiktionary's *responsiveness to neologisms and the level of detail and quality of definitions in its new word entries, when compared with expert-produced dictionaries?*

Objective 2 was addressed by identifying newspaper articles from across the research period containing selected neologism(s) from one of the three Dictionary Date of Entry datasets (referred to as 'stages' in the Objective (established in 3.4.4)). These new words were collected, along with key pieces of contextual information from the articles in which they appeared. Differing newspapers' use of these neologisms and new words' changing behaviour in the newspapers over the years enabled Research Question 2 to be explored:

> **Research Question 2 –** *What can be discovered about the 'neologic life-cycle' of selected neologisms in UK national newspapers between 2000 and 2014?*

Objective 3 built upon Objective 2, through the devising and piloting of the new manual methodology mentioned above, which created a 4.2 million word database of neologism-containing articles. In the longer term, the objective of this new methodology was to facilitate the collection of data for much larger context-rich, genre-specific corpora than has previously been the case. This would enable researchers to achieve the kind of deeply nuanced analysis of data that has previously only been possible only on a relatively small scale, for example, like the current study, examining the spread of neologisms in a particular genre. This new methodology represents a significant contribution of this project to academic study.

Objective 3 required a comparison of this new manual methodology with the most recently written-up example of the kind of automated webcrawling system which is currently used to extract data on neologisms (the *NeoCrawler*), in order to address Research Question 3:

> **Research Question 3 –** *In the context of data collection for context-rich, genre-specific web-based corpora, is the proposed new manual methodology more or less appropriate and effective in tracking neologism use and behaviour than the automated methods of the kind used by the* NeoCrawler*?*

4.2 Selecting Neologisms for Inclusion in the Study – *NeoCrawler*

In this section I present the neologisms selected for use in this study, and explain how this decision was arrived upon, including the required characteristics of candidate words and the spread of Word Formation Processes found within them. Thousands of new words are coined every year, filling lexical gaps often created by the advance of science or technology (the need to 'name' a new item or concept) or by events appearing in the media. Journalists are renowned for coming up with new words simply for entertainment value (for example 'Brangelina' to refer to the celebrity couple Brad Pitt and Angelina Jolie[58]). Other terms used to fill these lexical gaps can be 'catachrestic loanwords'. Catachrestic words are those which are generally considered to have been in some way misused, for example having been applied 'to a thing which it does not properly denote' (Oxford University Press 2016a). 'Catachrestic loanwords' fill lexical gaps which have been created by the introduction of a new idea or concept from the source language of the loan (Barrs 2015: 372).

The neologisms included in this study were those thought most likely to be of interest to fellow linguists and particularly researchers in the fields of neology and lexicography. The following factors were deemed to be of particular interest to members of these discourse communities:

- Word construction/formation
- Development and behaviour of new words:
    - Over time (14 years)
    - In response to external factors such as social or economic influences
- Differences in components of neologism entries in differing dictionary types, including expert-produced and collaborative

It was decided to start by choosing neologisms from an existing list of new words which had already been subject to monitoring and analysis, those generated by the

[58] See for example https://www.theguardian.com/film/2006/mar/19/features.angelinajolie

NeoCrawler program[59] (see 3.6.2). As these words had already been tracked over time within a specific genre, that of Google Blogs (Kerremans 2015: 80), it was felt that expanding this analysis into two new contexts (newspapers and lexicography) would provide a valuable additional layer of depth to both Kerremans' study and my own. This was something which could not be achieved by starting from scratch and identifying a new set of neologisms for study. At the same time, it would build upon the work of Renouf (2013) and her study of neologisms in *The Guardian* and the *Independent*.

The new words therefore chosen for this study were those shown in Table 4.1, organised by Dictionary Date of Entry Batches.

[59] http://www.NeoCrawler.de/crawler/html/

Neologism	Dictionary Date of Entry Batch
acedia (n)	DDEB3
bankster (n)	DDEB1+2
bogof* (n)	DDEB3
buzz marketing (n)	DDEB1+2
cold peace (n)	DDEB1+2
conurbation (n)	DDEB3
cyberbullying* (n)	DDEB1+2
cyberchondriac (n)	DDEB1+2
diabesity (n)	DDEB1+2
earworm (n)	DDEB3
e-tailer (n)	DDEB3
e-waste (n)	DDEB3
floordrobe (n)	DDEB1+2
frenemy (n)	DDEB3
gendercide (n)	DDEB1+2
globesity (n)	DDEB1+2
greenwashing* (n)	DDEB3
hubristic (adj)	DDEB3
hyperlocal* (adj)	DDEB1+2
newer markets	DDEB1+2
open education	DDEB1+2
predatory lending	DDEB1+2
promissory note (n)	DDEB3
rewilding* (n)	DDEB1+2
round pound	DDEB1+2
sodcasting	DDEB1+2
sovereign debt (n)	DDEB1+2
superphone* (n)	DDEB1+2
tablet computing	DDEB1+2
tenebrous (adj)	DDEB3
upskill (v)	DDEB3
warrantless (adj)	DDEB3
waterboarding* (n)	DDEB3
wellderly (n)	DDEB3

Table 4.1: Neologisms selected for this study

These new words were selected from the list of new words generated by the *NeoCrawler* project. It had originally identified and monitored 322 neologisms within the Google Blogs environment. For this study, it was decided that a maximum of 40 new words would be used. This would allow for detailed analysis of individual words, yet at

the same time enable patterns of behaviour to be observed. As this was an exploratory study, devising and utilising a new methodological approach, it would also lay the groundwork for further study in this field, using larger datasets identified and collected in the same way, and working in different genres where contextual information is equally important to the linguistic researcher. Thus it would be necessary to select 40 words from this much larger pool.

As one purpose of this research project was to examine neologism use and behaviour in newspapers, a key criterion in selecting suitable neologisms was that the word/phrase must have appeared in at least one of the project's newspapers during the qualifying period of the study: January 2000 – August 2014. This matched the period during which neologisms could have entered one or more of the project dictionaries. Neologisms which had appeared at least once in one or more of the project newspapers (*The Guardian*, the *Independent*, the *Mail*, the *Express* and *The Sun*) were accepted into the study. 'Media scoping' studies were therefore conducted in order to identify any neologisms which had not appeared in these newspapers; see 4.2.1. Before this could take place, however, it was necessary first to exclude several categories of words from the *NeoCrawler* list:

- Words without meaningful definitions. 'Meaningful' indicates that the definition is both complete and makes sense, thus the neologism 'chock' was excluded because its definition ('(techn. women's soccer): blend of') was unfinished (*NeoCrawler* list, Ludwig-Maximilians Universität n.d.)

- Trade names. Brand names for particular products, or the names of companies, for example 'iPod' in the neologism 'iPod league'. Also proper nouns, slogans or neologisms which are such 'niche' words that we would not expect to find them in normal everyday usage, for example 'dilscoop': 'a certain cricket move ("batting stroke")' (Ibid)

- Words which appeared to be neologisms but which were actually probably simple misspellings of existing words, or were catachrestic terms. For example 'loadly' – 'a lot of loudness' (Ibid) is most likely a misspelling of 'loudly'

- Words created using non-standard spelling conventions, for example the addition of an ‘a’ at the end of a word to add emphasis, as in ‘hella’ to mean ‘an intensive in Youthspeak, generally substituting for 'very', 'really', 'a lot'’ (Ibid)

- Non-British English words. Words either created in non-British versions of English, or words created for use in non-British English contexts were excluded, for example ‘teabonics’, meaning ‘new variations on English created by sign wielders at Tea Party protests’ (Ibid). (The Tea Party being an American political movement)[60]. Words with American spellings (such as ‘–ize’ endings) were not excluded, although in the event, none were selected for the study

- Terms made up of more than two words. This study confined itself to the examination of neologisms comprising one or two-word terms only.

Having excluded all of the *NeoCrawler* neologisms meeting these criteria, the list of potential new words for inclusion in the current study comprised 176 terms. These were then subjected to two ‘media scoping studies’.

4.2.1 Media Scoping

In this section I explain the media scoping process used to identify how many of the *NeoCrawler* neologisms actually appeared in the newspapers chosen for this study, and the difficulties presented by ‘false positives’.

As mentioned in 4.2, part of the neologism selection process comprised running a brief ‘media scoping’ study, in order to assess which words appeared in newspapers and which did not. Those which did not appear, were naturally excluded from the study, since new words with no media presence were of no value in a project examining how neologisms usage developed in newspapers over a 14 year period.

Although the intention was to run a single media scoping project, to determine how often each neologism was used in ‘online+print’ versions of newspapers. (The term ‘online+print’ indicates that the word had appeared in both print and online editions of the newspaper. This was to ensure that the articles being studied had potentially been

[60] See http://www.teaparty.org/

seen by all of the newspapers' readership, rather than one section or the other.) In practice, a second study was required, due to difficulties presented by false positives, the Google 'Right to be Forgotten' and *The Sun* newspaper.

In the initial media scoping study, the internal search engines of the five newspapers selected for the project were used, with the original intention being to continue using them during the full media tracking project. These newspapers were:

- The *Independent*
- *The Guardian*
- The *Mail*
- The *Express*
- *The Sun*

It was originally planned that only new words appearing at least 10 times across the five newspapers during that timeframe would be considered as potential neologisms for the study, since this would suggest that a word was beginning to becoming established in the news media. However in practice this resulted in the exclusion of many words which appeared otherwise likely to produce useful results. It was therefore decided that simply having appeared in one or more of the newspapers during the 14 years was evidence enough of potential media interest in the word. However 11 of the words which seemed to appear in the newspapers between one and 10 times were actually not appearing at all; instead, 'false positive' results were being returned (positive results erroneously returned by a search) for example 'iPadable' returned positive results which were actually for 'iPad').

These 11 words in fact did not appear in the newspapers at all. These false positives came to light as a result of extremely high numbers of media instances returned by *The Sun*'s internal search engine. Some of these numbered in the tens of thousands, for example 'half-false' (meaning 'quasi-false; not entirely true, not entirely false'

(*NeoCrawler* list, Ludwig-Maximilians Universität, n.d.)), returned 86,995 positive results in *The Sun*, and 'e-waste' ('electronic products which have been discarded of [sic] have become useless' (Ibid)) returned 12,832 results. It seemed unlikely that *The Sun* would use new words so much more often than any of the other newspapers, and indeed on re-testing using an external search engine (see below), 'e-waste', returned just 178 instances (129 of which were deemed unsuitable for collection using the new methodology outlined in 3.7). 'Half-false', meanwhile, returned no results at all once the external search engine had excluded all of the false positives (some returned in error for 'half' and some for 'false').

False positives were not the only problem encountered with internal search engines. There were also several inconsistencies identified across these newspapers, which made accurate comparisons difficult to achieve. These included:

- Dates – some newspapers only searched the last three years, others searched the entire online archive
- Presentation of results – for example if results were presented in date order, it was unclear whether this referred to the date of original publication or the date of the most recent Reader Comment (see 4.2.3), and whether this was the date shown on the search results page or not
- Content searched – some newspapers searched only staff articles, others included user-contributions and even advertisements
- Introduction of paywalls – unexpectedly limiting access to newspaper content
- Unexpected changes in search engine functionality – any of the above elements could change at any time, undermining the possibility of replicating the study at a future date

From this evidence, it appeared that the internal search engines were not discerning enough to be able to accurately identify new words. Although they were based upon standard commercial search engines such as Google, in each case the 'advanced' function was customised to the newspaper, and did not include all of the features of

the ‘advanced’ version of the commercial search engine. Not only did they not recognise Boolean Operators (which would have helped to exclude the false positives) they also did not offer the necessary range of search parameters, for example the ‘exclusions’ field.

It was therefore decided that the media scoping study should be repeated using external search engines, with their more targeted ‘advanced search parameters’, to try and establish a more accurate picture of neologism usage, particularly in *The Sun*. The search engine used for this task was Google Advanced Search. This was chosen simply because it is the search engine I am most accustomed to using, however as part of the development of the new methodology, a number of search engines were tested to identify the one which produced the most comprehensive and accurate results (see 4.3.1). This second media scoping study was indeed successful (as demonstrated above in the discussion of ‘e-waste’ and ‘half-false’). It excluded articles which did not contain the required neologisms, but caught all of the articles that did contain the new word (matching those returned by the internal search engines). It also searched the entire archive of articles online, rather than just those of a certain date or content type. (In any website, pages can be taken down and reposted at any time. This may be for maintenance or other purposes. Therefore the term ‘entire archive online’ should be taken to mean all of the articles available online at that moment in time. It is known that some articles have subsequently been removed and probably others have been posted which had been temporarily removed on the day of the collection. These changes are acknowledged but are ignored in conducting analysis of the database, since keeping up with the changes would require repeating the data collection on an infinite loop.) At the end of the second media scoping study, some 9,200 newspaper articles had been identified as containing the neologisms under study here.

4.2.2. The Sun vs Google Advanced Search

Here I outline the difficulties encountered when searching *The Sun* newspaper for instances of neologism usage, due to several webpages either disappearing and reappearing, or seeming to be inaccessible via Google.

Although the second media scoping study addressed most of the difficulties encountered when using newspapers' internal search engines, there was one area in which problems still remained. In the case of *The Sun*, some words/phrases that appeared in results pages from the internal search engine were not included in the external results. Similarly, words which appeared in the Google results on one day were sometimes found to be missing the next, and new ones had appeared in their place. Yet throughout, *The Sun*'s internal search engine confirmed that the 'missing' articles were present on the website, and they could indeed be accessed via the search results list. Consistently absent from the Google results were the terms 'floordrobe' and 'diabesity' (Creese 2015).

Initially, it was thought that the articles had been taken down, or that they were simply subject to normal website management, however their accessibility via the newspaper's own search engine suggested this was not the case. The difficulties already encountered with *The Sun* in terms of false positives (see 4.2.1) and the fact that the issue of 'missing' articles did not occur with any of the other newspapers suggested the problem was an internal one.

A possible alternative explanation was presented on hearing a *Radio 4* broadcast on the controversy surrounding the 'Google Right to be Forgotten' (RTBF), which entered law in May 2014 (Cox 2014). The new European Court of Justice legislation gives individuals the right to request that Google (and any other commercial search engine; Google is only associated with the ruling because it handles 90% of search requests in Europe (Ibid)) remove links to any story about themselves which they believe has become prejudicial (Preston 2014). This change forms part of European legislation on the processing of personal data (European Commission 2014: 1-2). It means that although the story becomes 'invisible' to Google, and hence will not be appear on any

search results page, it still exists on its home website and can be accessed via that site's internal search engine.

This would certainly explain the difficulties in accessing stories in *The Sun*, since their absence from the Google search results could simply be due to someone requesting that links to the story be removed, for any of the three established criteria:

- The story was inaccurate
- The information was no longer relevant
- The coverage was excessive

(EU Criteria for removal (Ibid))

While many of the requests involve cases of fraud, violent crime and child sexual abuse, others were not so clear-cut. Between May and September 2014, some 130,000 requests were submitted, approximately 50% of which were upheld without challenge, 30% were investigated and 20% rejected (Cox 2014). Articles may also be reinstated if the petitioner changes his/her mind, if Google changes its decision, or, presumably, if they have been removed in error (Lee 2014). It appears that now articles are being removed from Google indexing to comply with RTBF rules, a new possibility for erroneous removal of websites has been introduced into the system.

Since no reason is given on the website as to why its content has suddenly been blacklisted (Preston 2014), organisations have no idea whether the removal was legitimate or not, and this has led to major news outlets like the *BBC* (McIntosh 2015) and *The Telegraph* newspaper (Williams 2015) publishing lists of articles that have been removed under RTBF rules.

4.2.2.1 Impact of Google Right to be Forgotten on the Current Study

While investigation of these issues lies outside the purview of this research project, it is quite possible that the reason for the discrepancy between Google's Search Results Pages (SRPs) and those of *The Sun's* internal search engine could be that articles had

been removed from the former, either in response to an RTBF request, or simply in error. Their sudden reappearance could have been in an effort to correct such an error, or because Google had changed its mind on the original decision.

SRPs from which entries have been removed usually carry a sign-off line indicating that material has been removed for data protection purposes (see Figure 4.1). None of the Google SRPs for *The Sun* articles in this study carried this sign off, yet there is no way to prove that this in itself was not an error; certainly no other reason could be found for the removal and reinstatement of links to the articles under discussion, despite considerable effort to find alternative/additional explanations.

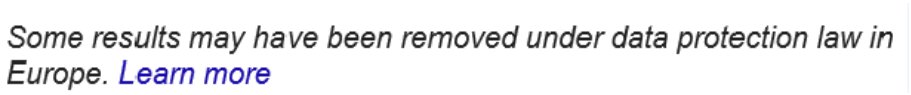


Figure 4.1: RTBF sign-off on Google search results pages

It was clear, however, that at any time the results of a Google search for neologisms in *The Sun* could produce completely different results to those achieved in a previous search. These inconsistencies between Google SRPs and those of *The Sun*, coupled with the thousands of false positives found in the newspaper could significantly cloud the results of this project. They would also make it very difficult, if not impossible, to replicate the study (see 3.3). As this is an exploratory study, devising, refining and testing a new methodological approach to corpus data collection, allowing for replicability (as far as is possible in any web-based project) was an important objective of the study. As a consequence of all this, the decision was taken to remove *The Sun* from the list of newspapers used for media tracking neologisms. This left the following neologism sources for analysis:

- *The Guardian*
- The *Independent*
- The *Mail*
- The *Express*

As these four newspapers still cover the socio-economic groups A, B, C1 and C2, D and E (see 3.5.1) this was considered acceptable, especially given that introducing a new paper at this stage would have required starting the neologism selection procedure from scratch, since media scoping was an integral part of it (see 4.2.1).

4.2.3 Identifying and Excluding Social Media Content – Reader Comments

In this section I discuss the way in which identification of a pattern to missing information in the Search Results Pages (SRPs) returned by Google represented one of the key proofs that a manual approach to the new methodology being developed here would likely result in a more streamlined approach to corpus data collection.

In a sample run of 20 search results to test potential upload procedures to a corpus query tool, it was found that in each case the neologism(s) actually occurred in the Reader Comments sections below the articles in the newspapers under study here, rather than in the articles themselves. None of the newspapers carried a feature for searching the Reader Comments section, and manually checking them was impractical since there were often upwards of 400 comments to a story.

However the same articles were noted on the SRPs to lack mention of the neologism in the text extract which appears below the hyperlink for each result. This is demonstrated in Figures 4.2 and 4.3.

Figure 4.2: Search result from *The Guardian* for the neologism 'earworm' – 'earworm' does not appear in the search extract

Figure 4.3: 'Standard' style search result from *The Guardian*, where the neologism 'earworm' does appear in the text extract

To investigate whether the two phenomena were related, SRPs for 'earworm' were examined, and 70 of the word's 178 *Guardian* search results did not include 'earworm' in the extract below the hyperlink. Several of those 70 articles were read in detail and it was found that the neologism did not appear there either. However on using the 'Find on this page' feature of the Internet Explorer web browser (see Figure 4.4), it was discovered that in each case, the neologism did appear in one of the Reader Comments below the article.

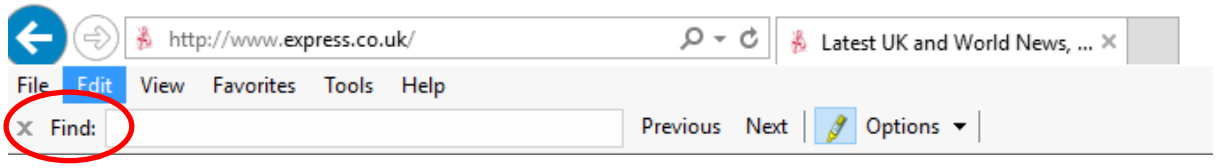


Figure 4.4: Internet Explorer's 'Find on this page' feature

To try to assess whether this was coincidental, or whether the lack of a search word in the SRP text extract meant that the neologism did not appear in the main newspaper article, a test was devised. Twenty URLs were checked to see if the neologism in question appeared in the SRP's text extract at the same time as appearing in the associated newspaper article. These 20 URLs were drawn from ten neologisms, two URLs for each neologism, some known to feature the search word in the extract and some not. URLs were selected from different newspapers, in order to check whether the issue affected only some or all of the publications. It was not possible to use a balanced sample of articles from each newspaper however, since not all newspapers featured all neologisms.

In each case, the newspaper article was read in detail, and the 'Find on this page' web browser function was used to locate instances of neologism usage. The results of this test can be seen in Table 4.2.

Neologism	**Newspaper**	**Neologism in GAS extract**	**Neologism in main article**
acedia	*Guardian*	N	N
acedia	*Independent*	Y	Y
bankster	*Guardian*	N	N
bankster	*Guardian*	N	N
cold peace	*Daily Express*	Y	Y
cold peace	*Guardian*	Y	Y
cyberbullying	*Guardian*	Y	Y
cyberbullying	*Daily Mail*	Y	Y
earworm	*Guardian*	N	N
earworm	*Daily Mail*	Y	Y
e-waste	*Guardian*	N	N
e-waste	*Guardian*	N	N
floordrobe	*Guardian*	Y	Y
floordrobe	*Independent*	Y	Y
frenemy	*Daily Express*	Y	Y
frenemy	*Guardian*	N	N
newer markets	*Independent*	Y	Y
newer markets	*Daily Express*	Y	Y
open education	*Independent*	N	N
open education	*Guardian*	Y	Y

Table 4.2: Results of test to determine whether absence of neologism from SRP text extract corresponded to absence from associated newspaper article

As can be seen from Table 4.2, there was a 100% correlation between neologisms appearing in the newspaper article and those included in the Google Advanced Search (GAS) (SRP) text extract. The converse was also true: absence of the neologism in the

text extract indicated absence from the article itself. This led to the hypothesis that if the search word did not appear in the GAS extract, the neologism would not appear in the newspaper article. Instead it would most likely appear in the Reader Comments section, although it could be anywhere on the page. (Since the only position of relevance to this study was in the main article, other locations were not considered relevant, although it was interesting that in all of the cases tested, the neologisms appeared in a Reader Comment.)

This hypothesis was double checked using a different set of 20 URLs. These were again identified by GAS, and each one contained no neologism in the text extract beneath the hyperlink. Again, two URLs from different newspapers were selected for each neologism. As shown in Table 4.3, none of the URLs was found to have a neologism in the main newspaper article. Instead, these were located in the Reader Comments sections.

Neologism	Newspaper	Neologism not in extract	Neologism in main article
bankster	*Guardian*	N	N
bankster	*Daily Mail*	N	N
cyberbullying	*Guardian*	N	N
cyber-bullying / cyber bullying	*Guardian*	N	N
earworm	*Guardian*	N	N
earworm	*Guardian*		N
green-washing	*Daily Mail*	N	N
greenwashing	*Guardian*	N	N
hubristic	*Guardian*	N	N
hubristic	*Guardian*	N	N
hyperlocal	*Guardian*	N	N
hyper-local / hyper local	*Daily Mail*	N	N
rewilding	*Guardian*	N	N
re-wilding	*Guardian*	N	N
sovereign debt	*Guardian*	N	N
sovereign debt	*Daily Mail*	N	N
superphone	*Guardian*	N	N
super-phone	*Daily Mail*	N	N
warrantless	*Guardian*	N	N
warrantless	*Guardian*	N	N

Table 4.3: Results of repeated test on SRPs and Reader Comments

This was taken as proof that if the search word did not appear in the GAS extract then the neologism did not appear in the newspaper article. Thus all of those search results noted to have no neologism in the text extract could simply be ignored when harvesting data (see 4.5.4). This significantly reduced the number of articles to be harvested, but would indeed require a manual approach since such exclusions could not be made using automated means. (It was noted, however, that where the neologism appeared in the first one or two of the Reader Comments, it did also appear in the SRP text extract. This was presumably due to some coding issues either with the newspaper or with Google. These fewer instances therefore had to be removed during the next phase of the data collection process.)

The discovery that elements of the search results themselves could be used to 'pre-exclude' articles that did not belong in the database/corpus supported my view that a new methodology would be most effective for the collection of data for context-rich, genre-specific corpora.

4.2.4 Identifying and Excluding Social Media Content – Blogs

In this section I discuss issues surrounding the exclusion of newspaper blogs – particularly blogs appearing in *The Guardian* – in order retain the integrity of this as a study of newspaper articles.

A decision was taken not to include uses of neologisms in newspaper blog posts, in part because the study was intended to collect standardised English written by professional journalists and in part because Kerremans (2015) had already addressed this in her original study of these words. However just as it was found that some newspaper articles did not contain the neologism in question in the article itself, but in the reader comments below, so it became apparent that there was the potential for confusion between neologisms appearing in newspaper articles and in their blogs. This related in particular to *The Guardian*, since it is the only one of the four newspapers to carry both its main website and its blogsite under the same internet domain name (guardian.co.uk). This presented major difficulties because it meant that poorly labelled blogs would be indistinguishable from articles. It would not be possible to rely on the language of the piece, since the relaxed style of language and the nature of the topics in *The Guardian* mean that an article could easily be mistaken for a blog, or vice versa.

It would have been useful to be able to check the *Google Blogs* index used by Kerremans to see how many *Guardian* articles were erroneously listed there, since Kerremans has included a number of 'blogs' in her study that were actually articles. For example the link and excerpt on page 103 of Kerremans' 2012 unpublished thesis, actually leads to a lifestyle feature and not a blog[61]. However, this was not possible. The database was accessed via http://blogsearch.google.com (Kerremans, Stegmayr

[61] https://www.theguardian.com/lifeandstyle/2010/feb/04/pregnant-women-forgetful-science

and Schmid 2012: 65); unfortunately this URL was deactivated in 2011 (Third Door Media 2014). While access to a limited number of blogs had been available for several years via *Google News* (and it had theoretically been possible to force Google to display the full database, although I had never found this to be successful) (Internet for Lawyers n.d.), in 2016 access was removed permanently (Ibid 2016). The *Google News* site has also now changed so that it is no longer possible to access blogs other than the official *Google News Blog* at the bottom of the page (see for example https://news.google.co.uk/). The best Google-related blog search now possible is a five-step process which provides access to a list of blogs which, it is claimed is still not as 'extensive or diverse as those retrieved by the previous BlogSearch' (Internet for Lawyers 2016).

As a result of these changes, I was never able to explore the full *Google Blogs* database used by Kerremans.

4.2.4.1 Categorising Guardian *Blogs and Articles*

Blogs are one of a number of social media tools that 'enable people to connect, communicate and collaborate' (Hemsley and Mason 2012: 3928-9). They allow people to produce and share knowledge across geographical boundaries (Ibid).

Articles with a lighter, less formal writing style than 'ordinary' *Guardian* pieces can appear to be blogs. Similarly, pieces written in the first person or on very personal topics (such as the article cited by Kerremans (2012: 103)) can be erroneously considered to be blogs, since they do not seem suited to a national newspaper. However comparison with articles from all sections of the hardcopy of the newspaper, particularly the Saturday and Sunday (*Observer*) editions, makes it clear that there is a much greater range of writing style in *The Guardian*. Articles regularly appear in these editions written in the first person, in a very informal even casual style, and concerning very personal issues such as the pressures of being a child prodigy[62].

[62] 'I could never live up to being a child prodigy', http://www.theguardian.com/lifeandstyle/2008/dec/20/family-child-prodigy.

This led to the establishment of the following criteria to pre-emptively identify and exclude blogs:

- Article has the word ‘blog’ in its URL
- Article has the word ‘blog’ in the title/headline of the page
- Article would not appear in the printed version of the newspaper, for example because it is regularly updated during the course of an event (for example a football match)
- Article contains reader contributions, for example screenshots of tweets in real time
- Article has a link at the bottom saying ‘More blogposts’, indicating that the article itself is a blog

There is no consistency in how a blog is identified however; any one of the features above could apply, and often the word ‘blog’ can appear either in the URL or on the actual page of the article, but not both (see for example Figures 4.5 and 4.6).

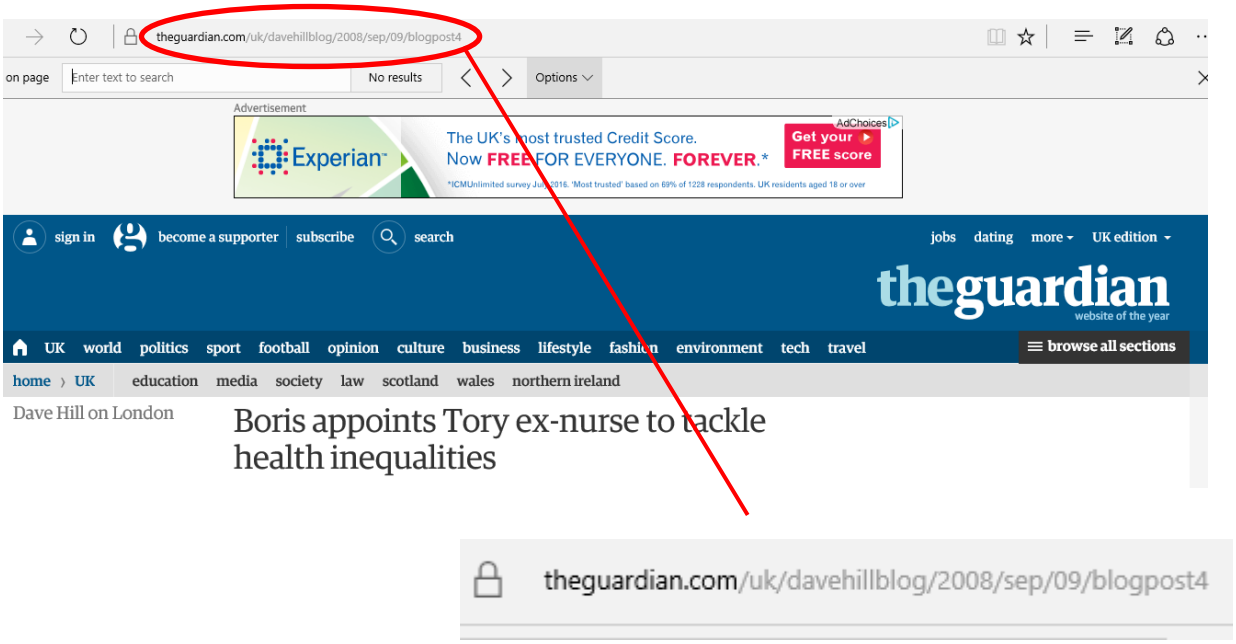


Figure 4.5: *Dave Hill Blog*, September 2008, with the only indicator being the URL, which features the word 'blog'[63]

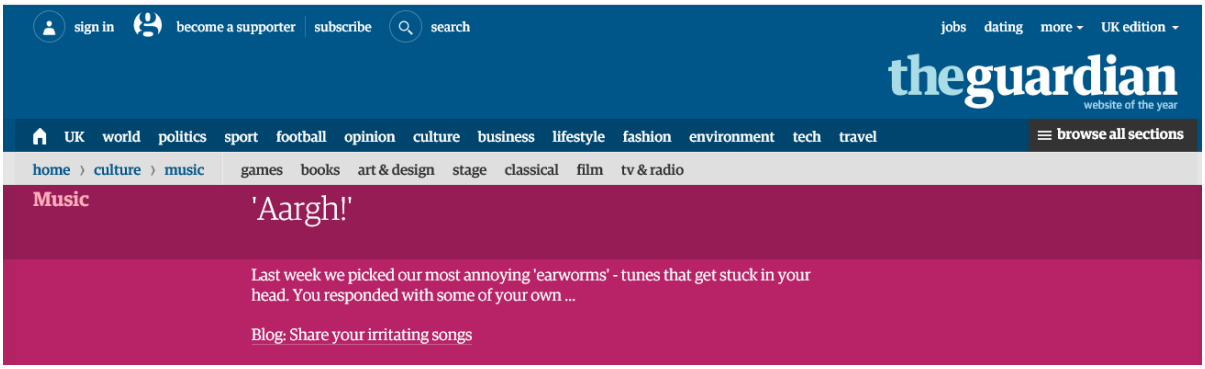


Figure 4.6: *Music Reader Blog*, June 2006, with the only indication being the word 'blog' at the bottom of the title[64]

Therefore any one of these characteristics is considered enough to exclude an article on the basis of it being a blog.

Through application of these criteria, it was possible to identify in excess of 30 blogs (see Appendix 2), most of which carry entries by *Guardian* staff, such as 'Datablog'[65]

[63] https://www.theguardian.com/uk/davehillblog/2008/sep/09/blogpost4
[64] https://www.theguardian.com/music/2006/jun/27/popandrock2

or 'Higher *Education Network* blog'[66]. However a few are by readers, for example for the *Music blog*[67]. Although largely written by professional journalists, all of these blogs were excluded from the study, since their purpose is to engage in dialogue with readers in a different way than the newspaper articles. It is also likely that different guidelines exist for staff on writing blogs, although due to the commercial nature of this information, such guidance is not publicly available.

It is possible that many of the 'Comment is Free' entries are submitted by readers, however since no indication is given of this, it is difficult to tell. In 2014 the *Guardian*'s 'Comment is Free' (CiF) opinion page was listed in its 'Help' pages as its main (which I take to mean flagship) blog, although it was never marked as such. It was launched as a 'group blog' in March 2006, and relaunched on a new platform in June 2008[68]. Having been identified as a blog, all articles with 'comment is free' either in the URL or the headline or title information were excluded from the study as above.

Today,' Comment is Free' is merely referred to as 'the home of *Guardian* and *Observer* comment and debate'[69], although it still appears in archives of blog posts, and is therefore still considered a blog for the purposes of this project[70].

4.2.5 Final Neologism Selection Process – Research Randomiser

In this section I show how the 96 remaining neologisms were whittled down into 40, using the Research Randomizer. Having conducted the media scoping studies, 96 neologisms remained. Most of the excluded terms had been rejected because they did not have usable definitions (110 words), or because they did not appear in any of the newspapers (83 words). The remaining 96 neologisms were checked against the five online dictionaries chosen for this study, to ensure that the definition provided by the

[65] See for example http://www.theguardian.com/news/datablog/2014/jan/09/cyberbullying-childline-statistics-online-bullying
[66] For example http://www.theguardian.com/higher-education-network/blog/2011/oct/25/open-access-higher-education
[67] http://www.theguardian.com/music/2006/jun/27/popandrock2
[68] See https://www.theguardian.com/commentisfree/2008/jun/04/1
[69] https://www.theguardian.com/help/2008/jun/03/1.
[70] See for example this article https://www.theguardian.com/commentisfree/live/2016/sep/15/should-grammar-schools-be-scrapped-join-our-debate-12-2pm-bst, accessible from this Blog Archive page: https://www.theguardian.com/tone/blog.

NeoCrawler matched that in at least one of them and where possible, to determine when words had entered the dictionary, in order that they could be assigned to a Dictionary Date of Entry Batch. Those five dictionaries were:

- *Oxford English Dictionary* (*OED*)[71]
- *Oxford Dictionary of English (ODE)*
- *Oxford Dictionaries* online *(ODO)*[72]
- *Merriam Webster*[73] *(MW)*
- *Wiktionary*[74]

Having checked each of the 96 neologisms against the five dictionaries, eight words were found to have *NeoCrawler* definitions which did not match those in any of the dictionaries

These eight words were therefore excluded from the project, leaving 88 *NeoCrawler* words from which to select the 40 required for my own study. This would be done through the use of 'randomising' software, specifically Research Randomizer[75]. This was used in order to avoid any possibility of bias being introduced into the study.

The 88 candidate words were split into their two groups based upon dictionary inclusion dates – 28 for January 2000 to August 2008 and 60 for September 2008 to August 2014 – from which 20 words would be selected for each group. For each date-category, the neologisms were input (in alphabetical order) into an Excel spreadsheet, in order to assign a number to each word. These numbers were input into the Randomizer in two runs (one for each date-group) and used to identify the words the program randomly selected.

[71] www.oed.com
[72] https://en.oxforddictionaries.com/
[73] www.merriam-webster.com
[74] www.wiktionary.org
[75] https://www.randomizer.org/

As shown in Figure 4.7, the user interface for the Research Randomizer is very straightforward.

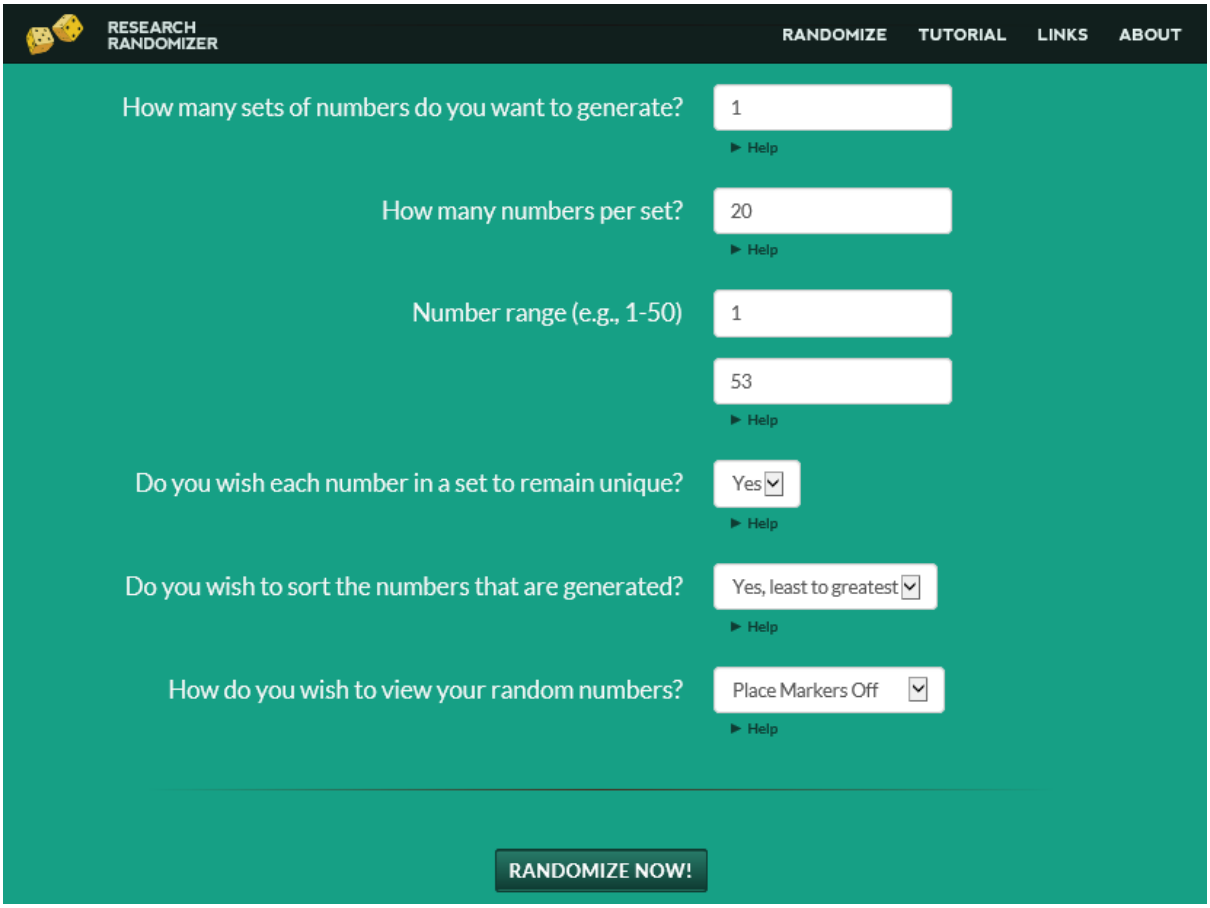


Figure 4.7: Research Randomizer user interface, completed in order to identify 20 neologisms from the 'date-group' September 2008 to August 2014, available at https://www.randomizer.org/

The Research Randomizer generates a series of numbers which can then be cross-referenced with the numbers in the Excel spreadsheet. The neologisms corresponding to the Randomizer numbers were thus selected as the neologisms for use in the final study. In this case, the program was run twice, once for 'date-group' January 2000 – August 2008 (Dictionary Date of Entry Batch 3 DDEB3) (selecting 20 neologisms from a possible 28) and once for 'date-group' September 2008 to August 2014 (DDEB1 and 2) (selecting 20 neologisms from a possible 60). The neologisms selected through this randomising process are shown in Table 4.1.

4.2.6 Adjustments to Neologism Lists

Here I discuss the final adjustments to the lists of neologisms produced by the Research Randomizer, and present the final list of new words for study.

The majority of the words returned by the Randomizer are nouns and this is unsurprising since, for example of the 28 better established words input into the Randomizer, 23 were nouns. Only three of the entire list were verbs. It was decided in the media tracking process to collect only instances of the inflection of the verb that appeared on the *NeoCrawler* list, rather than collecting every inflection, or choosing only to collect the infinitive of the verb. In practice, two of the three verbs included in the lists were in the infinitive ('liveblog' and 'upskill') and only 'sodcasting' was conjugated, in the present continuous form.

For nouns which have the same form as the present continuous form of a related verb, for example 'waterboarding', it was important that the database created here contain only the correct word class. In all but one case in this study, this was the noun (the exception being, again, the verb form 'sodcasting'). For each of the terms affected by this issue, where possible the wrong part of speech (the verb) was excluded from the study at the point of counting neologism instances per page.

These same issues arose in the second category of neologisms, those entering dictionaries from September 2008 to August 2014 (DDEB1 and 2).

These neologisms either entered the dictionary during the six-year period from September 2008, or had still not been accepted into a dictionary by the time data collection began in earnest (August 2014). As with those above, part of speech issues were addressed during the data collection and part-of-speech-checking stages of the project.

The original lists of *NeoCrawler* neologisms had included several terms which were variations in spelling of other words already on the list, for example 'under-share' as a variation of 'undershare'. The *NeoCrawler* searching mechanism had relegated these alternative spellings to a 'secondary results' category, suggesting that the system

considered them of less importance than the primary spellings. Only one version of each of these terms was included in the list of neologisms which was run through the randomising software. This was because there were several more terms on the list which also had potential spelling variants – consider for example 'rewilding', which can also be spelt 're-wilding' – yet these did not appear on the original *NeoCrawler* list. It had already been decided that these spelling variants would be included in data collection (see 4.3.2.1), and it was therefore unnecessary to include the second variant from the original source list as a separate neologism.

Having created the lists of neologisms for use in the study, it became apparent during the early stages of data harvesting that several of these new words were in fact not suitable for the current study, and would have to be removed. These terms are shown in Tables 4.4 (DDEB1) and 4.5 (DDEB3).

Neologism	**POS**	**Definition**
DDEB1 Neologisms not yet appearing in a dictionary as at 31 August 2014		
overparenting	N	overprotecting your children and preventing them from doing things independently out of worry

Table 4.4: Amendments to list of neologisms studied for DDEB1. (Definition source: *NeoCrawler* list, Ludwig-Maximilians Universität n.d.)

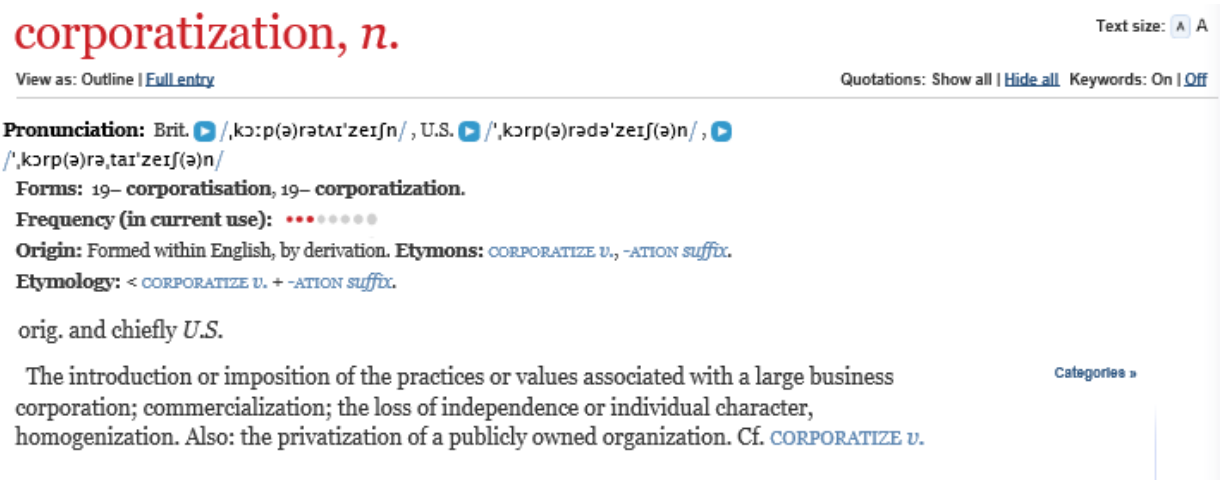
Overparenting. The term 'overparenting' presented grammatical problems, in the sense that most of the results returned for the term were actually phrases, where 'over' was used as a preposition. These were text extracts such as 'they argued over parenting', 'they had issues over parenting', and 'when he got home from work, he took over parenting duties'. None of the results were for the term 'overparenting' as defined by the *NeoCrawler* in Table 4.4, and consequently the term was removed from DDEB1. This left 19 entries for this list, 8 not yet in a dictionary and 11 already included. Similar difficulties would have been expected with any neologism formed in the same way, with a free morpheme used as an affix, such as 'over' or 'under'.

Further problems arose with several of the new words in DDEB3, as shown in Table 4.5, and again the decision was taken to remove them from the next stage of the project.

Neologism	POS	Definition
DDEB3 – Neologisms entering dictionaries between January 2000 and August 2008		
corporatization	N	the privatization of a publicly-owned organization
crackberry	N	nickname for the popular RIM communication device named Blackberry
ideation	N	the process of generating new ideas, where an idea is understood as the basic element of thought
liveblog	N	to write or update a blog at the same time as the event is happening
witricity	N	electronic process where energy is transferred without the use of wires

Table 4.5: Amendments to list of neologisms explored for DDEB3. (Definition source: *NeoCrawler* list, Ludwig-Maximilians Universität n.d.)

Corporatization. 'Corporatization' appears in all of the dictionaries in this study, however there are several different definitions or senses used (see for example the *OED* entry in Figure 4.8 below). (The US spelling of 'corporatization' has been used here, in line with Kerremans' usage. There were no further British-American spellings in the neologisms selected for this study.)



corporatization, *n.*

Text size: A A

View as: Outline | Full entry

Quotations: Show all | Hide all Keywords: On | Off

Pronunciation: Brit. /ˌkɔːp(ə)rətʌɪˈzeɪʃn/ , U.S. /ˈˌkɔrp(ə)rədəˈzeɪʃ(ə)n/ , /ˈˌkɔrp(ə)rəˌtaɪˈzeɪʃ(ə)n/

Forms: 19– **corporatisation**, 19– **corporatization**.

Frequency (in current use): •••○○○○

Origin: Formed within English, by derivation. **Etymons:** CORPORATIZE *v.*, -ATION *suffix*.

Etymology: < CORPORATIZE *v.* + -ATION *suffix*.

orig. and chiefly *U.S.*

The introduction or imposition of the practices or values associated with a large business corporation; commercialization; the loss of independence or individual character, homogenization. Also: the privatization of a publicly owned organization. Cf. CORPORATIZE *v.*

Categories »

Figure 4.8: *Oxford English Dictionary* entry for 'corporatization'[76]

[76] See http://www.oed.com/view/Entry/267635?redirectedFrom=corporatization#eid

The newspapers use all four of the *OED*'s senses, treating them as largely interchangeable. This made identifying appearances of the word using only the 'correct' definition (that included in *NeoCrawler* list: 'the privatization of a publicly-owned organisation') highly problematic. Rather than potentially skewing results with words which did not match the required definition, it was decided to remove 'corporatization' from the study.

Crackberry. A similar problem arose with 'crackberry'. Although the definition refers to the Blackberry device itself, matching *Wiktionary's* definition of the word[77], the media use of the term covers not only the device, but also users who are 'addicted' to their Blackberry, and the 'addiction' itself. At times it was difficult to be sure which meaning was intended, as demonstrated in Figure 4.9, and as with 'corporatization', this risked a skewing of results.

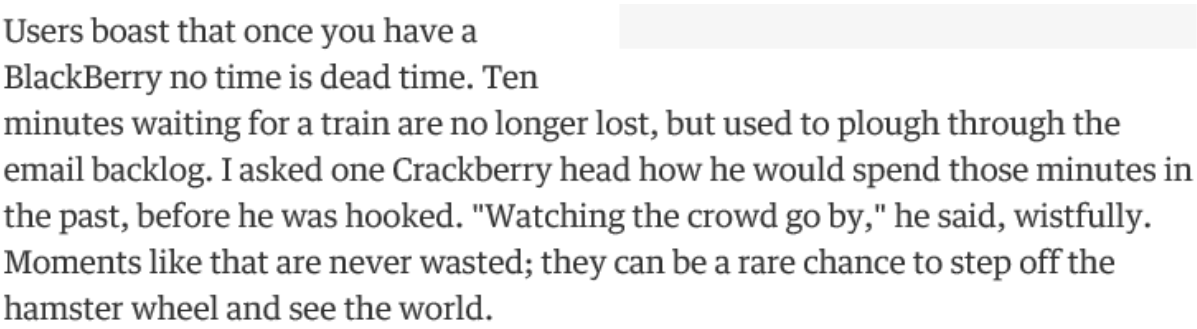

Users boast that once you have a BlackBerry no time is dead time. Ten minutes waiting for a train are no longer lost, but used to plough through the email backlog. I asked one Crackberry head how he would spend those minutes in the past, before he was hooked. "Watching the crowd go by," he said, wistfully. Moments like that are never wasted; they can be a rare chance to step off the hamster wheel and see the world.

Figure 4.9: Excerpt from *The Guardian* 'As a reformed addict, I can now see the full menace of a BlackBerry habit' by Jonathan Freedland, 22 August 2007[78]

It is unclear from the term 'Crackberry head' whether the author is referring to an addict (as in the common term 'crackhead' for someone addicted to crack cocaine) or a manager at the firm Blackberry. Either way, 'crackberry' was removed from the neologism list.

Ideation. The *NeoCrawler* definition of 'ideation' refers to the generation of new ideas, a theme reflected in the definitions of all the dictionaries used in the study. The use of the term in newspapers, however, tends to be confined only to one type of 'ideation',

[77] See https://en.wiktionary.org/wiki/crackberry
[78] See https://www.theguardian.com/commentisfree/2007/aug/22/comment.digitalmedia

mentioned only in the *Merriam-Webster* entry[79], that of 'suicidal ideation', or suicidal thoughts. These are defined by WebMD[80] as 'thoughts of ending a person's own life, or of killing one's self'. As the newspapers in this study were not returning results relating to the entire concept of 'ideation', but only one small aspect of it, it was decided to remove the term from the list of neologisms for the next stage of the study.

Liveblog. Thousands of entries for 'liveblog' were found in the newspapers, however during the actual data harvesting process (4.5 onwards), it became apparent that many of these were titles, for example for football match coverage[81].

Due to the difficulties already identified surrounding blogs in *The Guardian* (see 4.2.4), it had already become necessary to pre-exclude these from the data collection process. This meant that the neologism 'liveblog' could not be used because commercial search engines simply could not navigate the nuances of meaning involved in excluding blogs as an article category but retaining 'liveblog' as a neologism. There therefore remained no option but to remove 'liveblog' from the list of neologisms going forward.

Witricity. Witricity appears in the *NeoCrawler* neologism list as having two definitions:

- electronic process where energy is transferred without the use of wires
- the name of a company

The latter was excluded during the early stages of neologism selection as it is a proper noun. The first was retained, as it refers to the process rather than the company. However having completed data harvesting for 'witricity' it became clear that in fact these two meanings should really be a single definition. In every instance collected from the media, 'witricity' was not only the name of the process of transferring energy wirelessly, it was also the name of the concept/product as well. This was apparent by the fact that it was always spelled 'WiTricity' which, according to Alok Jha in *The*

[79] see http://www.merriam-webster.com/dictionary/ideation
[80] See http://www.emedicinehealth.com/suicidal_thoughts/article_em.htm
[81] See for example https://www.theguardian.com/football/2014/apr/18/football-league-clockwatch-live-mbm)

Guardian, is the name coined by the scientists at the Massachusetts Institute of Technology who developed it[82]. It is also the name of the company responsible for it[83].

Thus in addition to removing one neologism from the dataset for DDEB1 a further five were excluded from DDEB3. None were excluded from DDEB2 (neologisms entering *Wiktionary* and/or an expert-produced dictionary between September 2008 and August 2014). Tables 4.6 and 4.7 show the final neologism lists: 19 words for DDEB 1and 2 and 15 for DDEB3.

[82] See https://www.theguardian.com/science/2007/jun/08/wifi.gadgets.
[83] See http://www.independent.co.uk/life-style/gadgets-and-tech/features/a-world-without-cables-1822278.html

Neologism	POS	Definition
DDEB1+2 Neologisms entering dictionaries between September 2008 – August 2014, and new words not yet appearing in a dictionary as at 31 August 2014		
bankster	N	a person in the financial service industry who grows rich despite the continued impoverishment of those who depend on their services, and despite their apparent inability to succeed in business without constant government assistance
buzz marketing	N	word-of-mouth marketing
cold peace	N	strained political and diplomatic relationships between countries
cyberbullying	N	the use of Internet and mobile phones to send embarrassing or hurting [sic] messages
cyberchondriac	N	person who imagines they have a particular disease because their symptoms match those listed on an Internet health site
diabesity	N	diabetes caused by obesity
floordrobe	N	a very messy room where all the clothes are lying on the floor (the floor serving as wardrobe)
globesity	N	the idea that obesity has become a global problem
gendercide	N	systematic killing of members of a specific sex
hyperlocal	Adj	referring to the immediate surroundings; mainly used for referring to news
newer market	N	a region in the world where the production and import/export of goods is increasing
open education	N	educational organisations that seek to eliminate barriers to entry. Such institutions, for example, would not have academic admission requirements (e.g. distance learning programmes)
predatory lending	N	deceptive, fraudulent or abusive lending practices
rewilding	N	the process of returning species, habitats and landscapes to a natural state, as they would be without the intervention of humans
round pound	N	a price in whole pounds rather than a combination of pounds and pence; a selling strategy
sodcasting	V	to play music through the speaker on a mobile phone, usually on public transport.
sovereign debt	N	a debt instrument guaranteed by a government; a bond
super phone	N	smartphones with better performance, desktop-grade web browsing, and high-resolution displays
tablet computing	N	a style of computer technology which uses a tablet or slate computer (=a wireless computer with a touch pad)

Table 4.6: DDEB1+2 Neologisms not yet appearing in dictionaries, and neologisms entering dictionaries. (Definitions source: *NeoCrawler* list, Ludwig-Maximilians Universität n.d.)

Neologism	**POS**	**Definition**
DDEB3 Neologisms entering dictionaries between January 2000 and August 2008		
acedia	N	spiritual or mental sloth
bogof	N	an advertising strategy that entices people to buy a product and get one for free
conurbation	N	an extensive urban area resulting from the expansion of several cities or towns so that they coalesce but usually retain their separate identities
e-tailer	N	a company which uses the Internet to sell its products
e-waste	N	electronic products which have been discarded or have become useless
earworm	N	a piece of music that sticks in a person's head
frenemy	N	a person you assume as a friend, although you don't really like him/her
hubristic	Adj	referring to someone or something behaving with hubris
greenwashing	N	the practice of making an unsubstantiated or misleading claim about the environmental benefits of a product, service, technology or company practice
promissory notes	N	a negotiable instrument, wherein one party (the maker or issuer) makes an unconditional promise in writing to pay a determinate sum of money to the other (the payee), either at a fixed or determinable future time or on demand of the payee, under spec
tenebrous	Adj	dark and gloomy
upskill	V	to give employees extra training in order to improve their performance
warrantless	Adj	without a warrant; especially referring to governments' surveillance practices after 9/11
waterboarding	N	a torture method of putting a cloth over the face and pouring water over it to make them believe they are drowning
wellderly	N	old people who are in good health

Table 4.7: DDEB3 Neologisms. (Definition source: *NeoCrawler* list, Ludwig-Maximilians Universität n.d.)

4.3 Media tracking: Corpus Building and New Methodology

In this section I lay out the next steps in devising the new methodology to be used to collect corpus data for this project, specifically testing to identify the most suitable external search engine to use, locating articles containing the neologisms which had been selected for this project, and exploring the problems raised by spelling variants among these words.

Among the most important elements in the new methodology developed during this project were the methods devised for what I term 'data harvesting', a process comprising three key activities:

1. Identifying potential articles and then limiting the list of these to only those deemed suitable for inclusion in the corpus/database

2. Accessing the suitable newspaper articles, including key contextual information such as publication date

3. Collecting and uploading those texts into a corpus query program such as Sketch Engine (used in this study).

Before any of the tasks mentioned above could be undertaken, it was first necessary to identify the most appropriate 'external' search engine for use in the study.

4.3.1 Testing Commercial Search Engines

As mentioned in 4.2.1, having determined that internal search engines were not reliable or consistent enough to be used during the 'media tracking' phase of this project (the process of tracing the use of a set of neologisms in UK national newspapers online), it was necessary to select a commercially available 'external' search engine to use instead. This search engine would be used to complete Task One above, and the search results it produced would need to be compatible with software/methods chosen for Tasks Two and Three.

The search engines selected for comparison were Google, Yahoo, and Yahoo News. In each case, it was the 'Advanced Search' version of the search engine that was tested,

since the data harvesting process would require that searches be limited to specific internet domains (for example www.theguardian.com), and it would be necessary to search for exact words or phrases, and to have the facility to exclude certain terms such as false positives. Ordinary versions of search engines do not offer this level of functionality. It had been planned to also test the Microsoft Bing search engine, but I was unable to find an Advanced Search option for it. (It seems I was not alone, judging by the number of search queries I found asking for help finding it, returned when I ran a Google search for 'Bing advanced search').

Each of the remaining three search engines was tested to see whether accurate results could be generated through the use of 'exact phrase' and 'exclude phrase' fields on the Advanced Search Query Form (shown for each search engine at Appendix 3). The objective was to see if any search engine produced sufficiently nuanced results, distinguishing between the neologism (which was required) and false positives (which were not). The neologisms used for this test are shown in Table 4.8. (The 'error words' for 'cyber' are just two examples of the many false positives returned for 'cyberbullying', some based on 'cyber', some on 'bullying'.)

Required Neologism	**'Error word' leading to false positive result**
'iPadable'	'iPad'
'bankster'	'bank'
'gendercide'	'gender'
'diabesity'	'obesity'
'cyberbullying'	'cyber-harassment', 'online bullying'

Table 4.8: Neologisms and false positives generated by search engines

All of the search engines returned the false positive results as listed above. In each case, this was verified by manual reading of the sample articles chosen for testing, to check that they did contain the 'error' word and did not contain the neologism.

When the error words were entered into the 'exclusions' field on the search query forms, the following results were achieved:

'iPadable' / 'iPad'

- Yahoo Advanced Search – continued to return false positives for 'iPadable' and failed to exclude articles containing the term 'iPad'
- Yahoo News Advanced Search – as per Yahoo Advanced Search
- Google Advanced Search – returned no further false positives and correctly excluded articles containing the term 'iPad'

'Bankster' / 'Bank'

- Yahoo Advanced Search – successfully excluded 'bank', however also excluded any term containing the word 'bank' – 'banking', 'banker' and 'bankster'. The search word was therefore erroneously excluded as well
- Yahoo News Advanced Search – as per Yahoo Advanced Search
- Google Advanced Search – successfully excluded 'bank', but retained any terms containing the word 'bank' – the search word was therefore correctly returned as a search result

'Gendercide' / 'Gender'

- Yahoo Advanced Search – successfully excluded 'gender', but also excluded any term containing 'gender' – 'genders' and 'gendercide'. The search word was therefore excluded as well
- Yahoo News Advanced Search – as per Yahoo Advanced Search

- Google Advanced Search – successfully excluded ‘gender’, but retained terms containing ‘gender’, specifically ‘gendercide’

‘Diabesity’ / ‘Obesity’

- Yahoo Advanced Search – successfully excluded ‘obesity’, but also excluded ‘diabesity’. The search word was therefore excluded as well as the ‘error word’

- Yahoo News Advanced Search – as per Yahoo Advanced Search

- Google Advanced Search – successfully excluded ‘obesity’, but correctly retained ‘diabesity’

In all cases, then, Google Advanced Search was the only one of the three which correctly retained the search word, whilst excluding the ‘error word’. In a normal corpus, it would be considered problematic for a search engine to exclude a high frequency word like ‘bank’ (which has a frequency of 216.1 per million (normalised figure) in the British National Corpus (BNC) (accessed via Sketch Engine[84])). However in this case, the key word was ‘bankster’ not ‘bank’ (with a frequency in the BNC of zero), and therefore the measure of success for the tested search engines was whether or not they were able to distinguish between the two, and retain the former despite excluding the latter. This Google Advanced Search achieved.

‘Cyberbullying’ was more complicated than the other four terms, since it generated so many false positives, some replacing the word ‘cyber’ to talk about other kinds of bullying, and others replacing ‘bullying’ to talk about other kinds of ‘cyber attack’. The only way to address these false positives was to include a string of words in the ‘exclude’ field of the search form, for example ‘harassment’ (to exclude ‘cyber-harassment’) and ‘online’ (to exclude ‘online bullying’). Once again, only the Google Advanced Search Engine was able to cope with such complicated search parameters; it successfully made all the necessary exclusions and returned a set of results free of false positives.

[84] https://the.sketchengine.co.uk/

Out of all of the test runs then, Google Advanced Search (GAS) was consistently capable of dealing with the demands of this kind of data collection. The only area in which it was not the best choice lay in the fact that Yahoo News Advanced Search offers the possibility of limiting a search to a particular date range, which GAS does not. This would have enabled me to search specifically between 1 January 2000 and 31 August 2014. However the inconvenience of having to manually exclude article results outside of this date range was far outweighed by the advantage offered by more reliable and accurate results offered by GAS, and hence the latter as chosen as the search engine for use in this study.

4.3.2 Locating Neologisms for Data Collection

Having identified the 34 neologisms to be used in the study, and selected the most effective and accurate search engine to use, the search for neologisms within the four UK national newspapers could begin in line with the tasks laid out in 4.4. Here I discuss the automated functionality of the chosen search engine, Google Advanced Search that could have been employed in this project, to serve as a basis for comparison with later manual methods.

As mentioned in 4.2.1 by the end of the second media scoping study 9,200 articles had been identified as potentially containing neologisms. One of the first things to be investigated were potential methods of excluding large numbers of unwanted articles, however failures in the way in which Google Advanced Search operated supported my view that to do this before data collection, rather than after having to collect, assess and then exclude them in post-processing would be more effective. This represented a crucial stage in the development of the new methodology, demonstrating, as it did, that manual methods could actually be more suitable to data collection in a digital context than automated ones.

This began with a straightforward advanced search of the newspaper internet domain name for each neologism, using either Internet Explorer Version 7 or 8 (depending on the computer hardware available). For each new word, an initial Google Advanced Search (GAS) was run, specifying the exact neologism (in speech marks) in the ‘this

exact word or phrase' field of the search form, setting the language as English and restricting the search to the specific newspaper, for example www.theguardian.com (see Figure 4.10, using the sample word 'conurbation'), 'conurbation' is one of the words in this study which I term 'reincarnated', having fallen out of favour and later returned to mainstream usage (see 5.4.1.2).

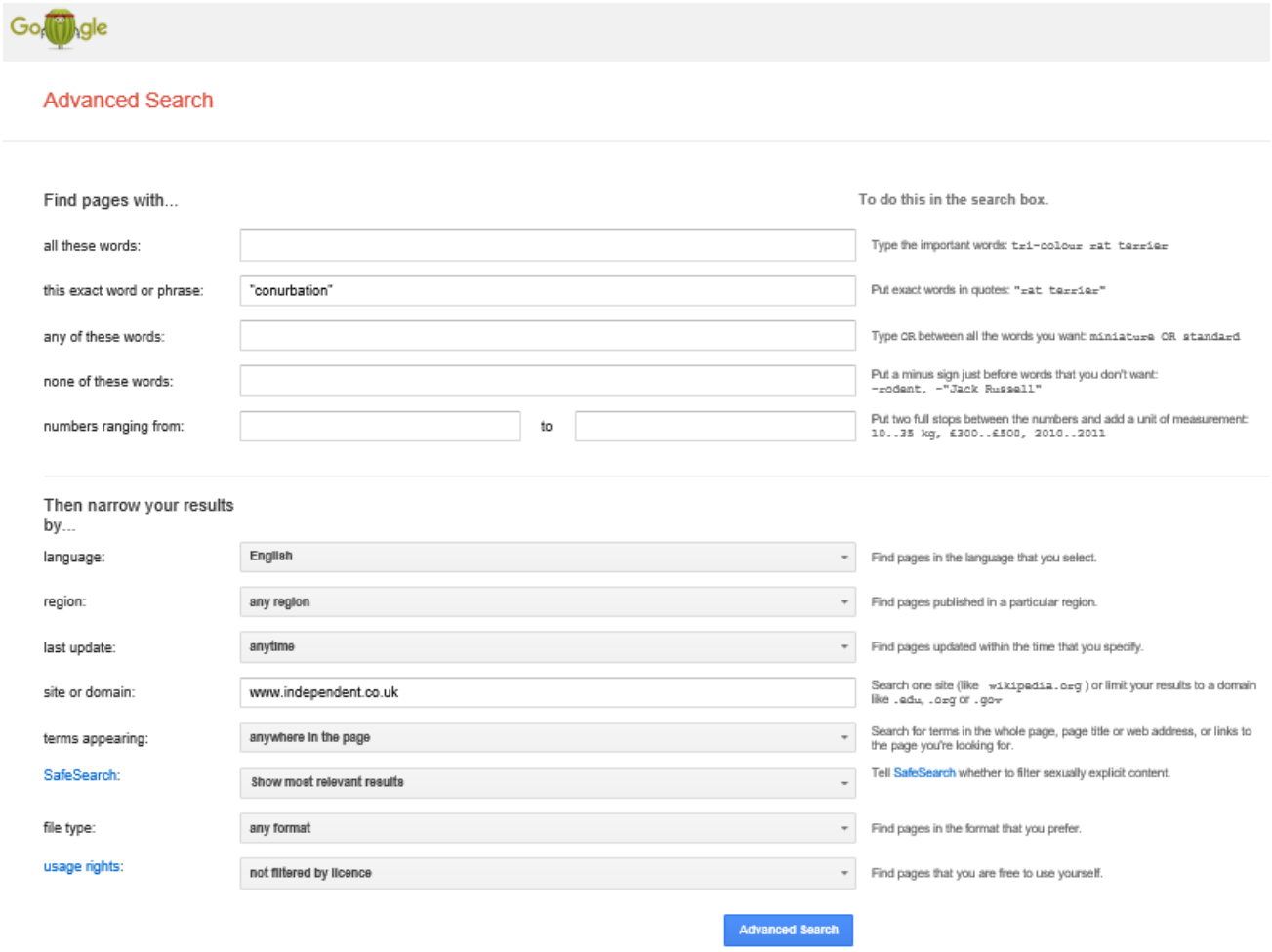


Figure 4.10: Google Advanced Search form as it appears on 14 August 2016 (in Internet Explorer 11, within Windows 10)

The 'terms appearing' field (fourth from the bottom) should have ensured that all results returned were for neologisms appearing in the main article on the webpage (using the 'in the text of the page' option) (and should therefore have avoided issues such as search words appearing in links – see below) (see Figure 4.11).

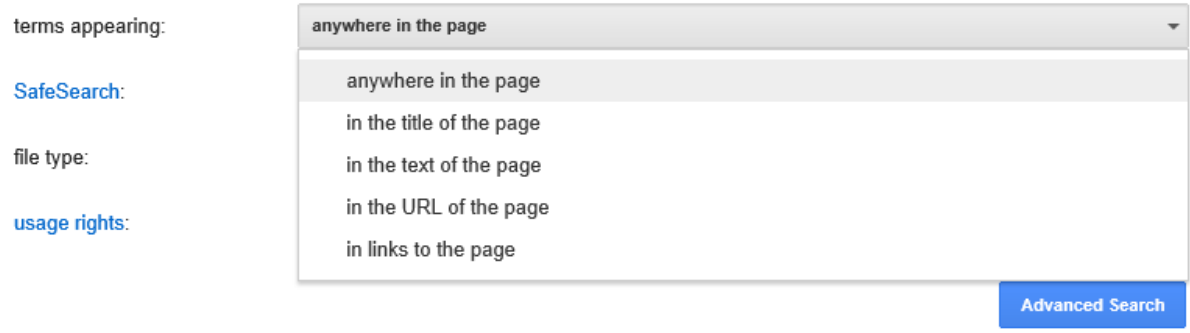


Figure 4.11: 'Terms appearing' drop-down menu

However this 'terms appearing' option was found to be unreliable. Several times during test runs it was found to have missed articles which a search conducted without using this feature did locate. It was therefore necessary to set this option to 'anywhere on the page', and manually 'pre-exclude' neologism uses not in the main article.

Using GAS returned lists of Search Results Pages (SRPs), each containing 10 results showing the title of the newspaper article, a hyperlink leading to the webpage, with a truncated version of the article's web address below (see 4.4.2 for discussion of truncated web addresses), the date of the article (although this was not always accurate) and an extract from the article showing the neologism in use (see Figure 4.12).

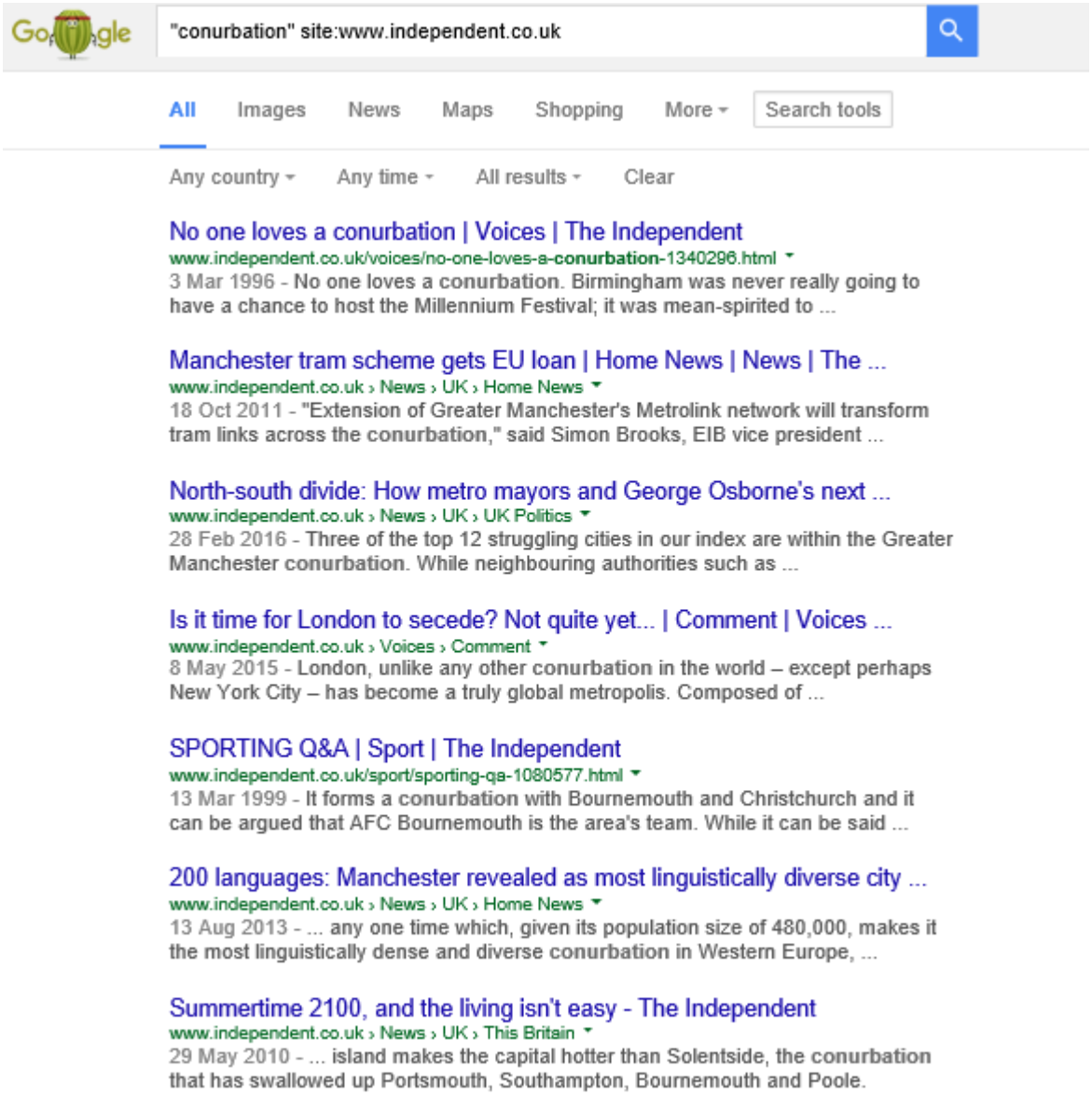


Figure 4.12: Search Results Page for 'conurbation' in the *Independent*, collected 14 August 2016 (collected by Internet Explorer 11, within Windows 10)

Yet the results on this SRP are not presented in date, alphabetical or any other discernible order. This is because results appear on SRPs according to Search Engine Optimisation (SEO), a procedure whereby webmasters code their sites to try to provide users with as many results as possible. These not only fit the user's ideal search criteria, but also what the webmasters hope will be useful tangential information or formats as well. It is generally assumed that users will want as much information as possible, and so sites are coded in such a way that the algorithms used by search engines will choose an individual page in response to many seemingly unrelated searches (Evans 2007: 21-22). As this particular search was conducted in August 2016, there were likely to be

many entries for 2015 and 2016 (especially since there are known to have been 299 articles featuring 327 uses of the word between 2000 and 2014) however because the results are not in date order, only one of these newer entries is visible on this SRP (third entry from the top). At the time of data collection for this study, there were already regular instances of articles published after the study's end-date for dictionary and newspaper inclusion (31 August 2014) and these, and articles published before the start date of 1 January 2000 were dealt with as part of the next stage of project: URL harvesting (see 4.4).

4.3.2.1 Spelling Variants

Here I lay out the neologisms in the study which were at times spelt in different ways, either as a compound form, a hyphenated form or a two-word term.

As mentioned in 4.2.6, there were a number of words included in the neologism list used for this study which could be spelt in several different ways. All bar one of these could be spelt either as a compound/blend, for example 'greenwashing' (meaning 'the practice of making an unsubstantiated or misleading claim about the environmental benefits of a product, service, technology or company practice' (*NeoCrawler* list, Ludwig-Maximilians Universität n.d.)), as a hyphenate 'green-washing' or as a two-word term 'green washing'. 'Re-wilding' was found to be spelt either as a hyphenate or a compound; no two-word spellings were found.

When we examine the list of neologisms under study here, we find that the majority of the new words cannot be broken into hyphenated or separate parts, since they feature 'bound morphemes' which cannot function alone, for example 'gl' in 'globesity', 'dia' in 'diabesity' and 'ster' in 'bankster' (Katamba 1994: 54-5). However some bound morphemes appear in hyphenated rather than compound form, for example in this study, 'cyber-bullying'. Test searches were run in the newspapers on 'cyberbullying' and 'cyberchondriac' and it was found that the former could be spelt in any of the three ways ('cyberbullying', 'cyber-bullying' or 'cyber bullying'), yet the latter only ever appeared as a single unhyphenated word. Further tests were run on each of the words

that might feasibly break at the 'joining point' of the compound/blend; the results are shown in Table 4.9.

Neologism Compound/Blend Form	**Neologism Hyphenate/Two-Word Forms**
cyberbullying	cyber-bullying/cyber bullying
cyberchondriac	*
earworm	ear-worm/ear worm
greenwashing	green-washing/green washing
hyperlocal	hyper-local/hyper local
rewilding	rewilding**
sodcasting	sod-casting/sod casting
superphone	super-phone/super phone
upskill	***
waterboarding	water-boarding/water boarding

Table 4.9: Neologisms with potential spelling variants
* No results were found for 'cyber-chondriac'/'cyber chondriac'
** 'Re' is a bound morpheme. Although it is commonly hyphenated, as in 're-wilding', it does not take the two-word form.
*** Only compound forms were found for 'upskill'; no variants were present in any of the newspapers

Kerremans had reported that when she searched for words with potential variant spellings, Google returned all three in her search for any single one (2012: 55), however my experience did not correspond with this. My testing showed that searching for a hyphenated or two-word term did indeed return results for both forms. However it did not return results for the compound/blend. Similarly, a search for the compound/blend returned only one-word results.

It was therefore decided that two forms would be searched for, for each of the neologisms with potential spelling variants (as shown in Table 4.9). These would be the compound form, and the hyphenated form, the latter of which would also return results for the two-word form, as had already been shown.

This was considered particularly important in this study on newspapers since during testing it was found that the same newspaper would regularly use any of the three spelling variants, often in the same article[85].

4.4 Automated Methods of URL Harvesting

In this section I present the various automated systems for downloading Google Advanced Search results which I tested in order to be certain that the planned manual methodology was indeed the way forward. Having produced 9,200 initial search results for the neologisms in all four of the newspapers in the study, the next step was to download each of these articles and access the contextual information contained within each one. Mindful of Lüdeling, Evert and Baroni's assertion that 'the process of downloading webpages to build the corpus … must be automated) (2007:25), I first set out to find and test the automated methods that would normally be used to take this project on to the next stage, before developing the planned manual approach. This would clearly be a process of trial and error, and it quickly became apparent that in order to gather all the information I required for my study the process would need to be broken into two separate components: harvesting of URLs, followed by the collection of the following contextual information:

- newspaper title
- publication date
- number of instances of neologism in the article
- whether neologisms appeared in the headline
- article type (for example 'news')

[85] See for example http://www.dailymail.co.uk/news/article-2441239/1-5-young-people-suffer-extreme-cyber-bullying-day-Facebook-accounting-half.html, which uses 'cyber-bullying' in the headline and 'cyberbullying' in the text. This is likely because headlines are written by sub-editors who lay out pages, while articles are written by journalists (based on my own personal experience as a journalist and sub-editor)

I began with data harvesting, investigating the following methods of downloading search results:

- designing bespoke data collection software
- repurposing data management software
- programming new commands into Dragon voice-activated software

4.4.1 Bespoke Corpus Data Collection Software

Given the genre-specific nature of this project, initial investigations were made into the possibility of creating some kind of bespoke data collection software to follow the hyperlinks on each Search Results Page (SRP) and then copy and download all of the contents of the target web pages either into Microsoft Word or directly into a corpus query program such as Sketch Engine, depending on how the next step – collection of contextual information – was to be carried out.

Several experts were approached about the feasibility of creating an automated URL downloading program, including Adam Kilgarriff of LexMasterClass (creator of Sketch Engine), Alex Bennett of Portsmouth University (who had recently completed a new web-as-corpus creation tool with some similarities to this project) and Professor Zhang Yihua of Guangdong University of Foreign Studies, in China. Professor Zhang had, with the help of a student, created a similar programme to gather data from another major media outlet. None of these experts thought it would be possible to download from Google SRPs, or directly from the newspaper websites. Bennett argued that each webpage, even within the same website, is coded differently, making it almost impossible to write a programme capable of searching every individual page (Personal Communication, June 2014). Zhang pointed out that access would be required to the site's entire internet domain (a closed system), and that this was simply not available for sites like newspapers (Personal Communications, May-July 2014). Kilgarriff felt a bespoke solution was simply unfeasible (Personal Communication August 2014).

Having rejected the idea of a search-result-downloading tool, I moved on to the possibility of repurposing other types of software designed to work on/with the hyperlinks appearing on the Google SRPs.

4.4.2 Repurposing Data Management Software

One potential option was the repurposing of existing data management software, for use as data collection software. I conducted considerable research into this idea, including:

- Trying to change the settings on the presentation of commercial search engine results, to display full URLS rather than just parts of a URL below the hyperlink. This would allow the web addresses to be copied directly into Microsoft Word, or to be uploaded directly into a corpus query program without actually having to clicking on them, open the target webpage and select the URL from the web address bar itself.

- Looking for software to:

    - Restore truncated URLs on SRPs, for the same reason. These URLs are always truncated – showing just the beginning and the end of the web address – so that they cannot simply be copied from the SRP itself. This may be an 'anti-bot' measure, designed to prevent automated programs (such as the one I was trying to create), or 'bots' (similar 'unmanned' programs) from downloading results in bulk. This is probably because such 'bots' are used to gather web addresses to then either spam or in some other way interrupt normal business practices.

    - Re-purpose software programs designed to:

        - Repair broken links, which might potentially provide the full URLs for the hyperlinks on SRPs

        - Retrieve files such as the newspaper article webpages, for example using 'Wget'[86]

    - Automatically follow all the hyperlinks on an SRP

    - Find alternative ways to collect URLs in bulk, using the advanced functions of alternative search engines such as KISSmetrics[87] and ixquick[88]. Both of these are advanced searching and analysis tools, which were investigated but deemed unsuitable since they are designed for commercial rather than academic use. KISSmetrics is a marketing device designed to improve customer experience, while ixquick provides maximum search results from multiple search engines, meaning that it is not refined enough to be of use in this project.

None of these options, proved viable. The only piece of software that was considered potentially useful was from Internet Marketing Ninjas[89], and was a programme designed to extract URLs from search results pages[90]. Use of this programme required the downloading of two separate applications; although downloading of the first was successful, accessing the second was not. Despite numerous attempts on several different computers using several different web browsers and multiple internet service providers, it was never possible to access the second of the two applications. This avenue was therefore reluctantly abandoned.

4.4.3 Computer-Aided Data Harvesting: Voice-Activated Software

In order to speed up the process of manually clicking and collecting the URL from each Google search result, I used Dragon Naturally Speaking voice-activated software (Nuance 2014), although as with all of the previous potential solutions, this still left unanswered the question of collecting contextual information. As previously

[86] See http://www.gnu.org/software/wget/
[87] See https://www.kissmetrics.com/
[88] see https://www.ixquick.com/
[89] See https://www.internetmarketingninjas.com/
[90] See https://www.internetmarketingninjas.com/seo-tools/get-urls-grease/

mentioned, however, it was decided to try and resolve the article-accessing problem first, and to move on to the issue of contextual information afterwards.

To use Dragon in this context, I wrote coding to allow URLs to be automatically collected in batches of 10 (the maximum number of search results on a page), in response to a single verbal command (see Figure 4.13).

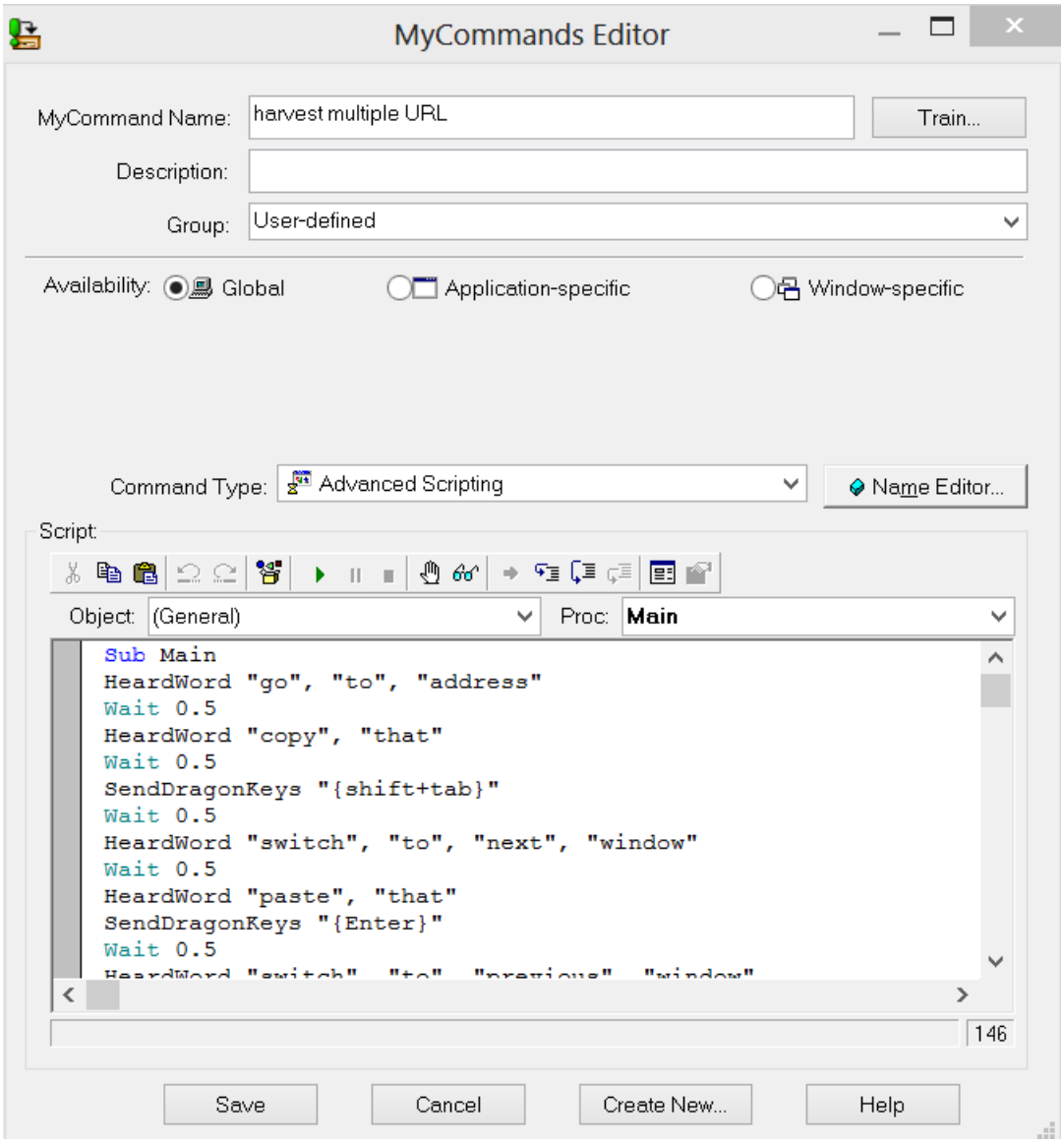


Figure 4.13: Multiple URL harvesting command, written by myself and added to Dragon Naturally Speaking V12.0 Command Browser (Nuance 2014)

The 'multiple URL harvesting command' (triggered by the voice command 'harvest multiple URL') meant that all 10 articles on an SRP for a particular neologism could be collected in a row, without further user intervention. The program would cause the

mouse (which was active in Internet Explorer) to click on the first hyperlink on the SRP, then when the webpage opened, it would highlight the web address line and copy the URL. It would then switch to Microsoft Word, where it would paste the URL, then switch back to Internet Explorer to select the next result's hyperlink, and repeat the process. This would be done 10 times, before the program would stop and await further instructions. Given enough computing power, it would be possible to further the program, so that it then clicked on 'next' at the bottom of the page, opening up a new SRP and beginning the process all over again. However the limitations of computing power currently available meant that not only was this further development not possible, but the current program was highly unstable, leading to regular software crashes due to the complexities of the automated task. This meant that for the time being, it was quicker and more reliable to harvest the URLs by hand, although with the necessary computing power it will be possible to use similar Dragon programs for future research of this type, and it may even be possible to programme Dragon to locate and collect contextual information.

4.5 The New Manual Methodology of Corpus Data Collection

In this section I explain in detail the new methodology used in this study to create the *NTON* database (*Neologism Tracking in Online Newspapers*), including methods of 'pre-screening' 'pre-exclusion' and 'advance exploration' of websites as means to vastly narrow down the number of corpus texts being collected. Having investigated all of the potential automated methods of collecting data identified by Google Advanced Searches, and found each to be wanting, it was clear to me that, despite working in a digital, web-based context, the most appropriate means of accessing these search results would be to apply manual methods. A manual approach might also be used to:

- reduce the amount of post-processing required once the database (or, in future corpus) was complete, before analysis could begin, and
- facilitate collection of more contextual information, allowing for the building of a more targeted and larger context-rich database/corpus in a particular genre,

where previously the limitations of automated methods – for example the inability to collect accurate date information (see 3.7.2) – had constrained corpus size and complexity.

Media usage for most of the neologisms discussed in this thesis, particularly those in DDEB3, probably began before 2000, although the *NeoCrawler* treated them as new words. This is one of the difficulties with the *NeoCrawler's* automated system, since it was designed to collect only words which had not been used before 2006. In particular, the *NeoCrawler* identified as neologisms four words which first came into use decades ago. However they dropped out of use and then recurred, leading me to refer to them 'reincarnated' terms (see 5.4.1.2). Since these words had been identified as 'neologisms' by the *NeoCrawler*, and I am comparing my new methodology with its existing automated systems, I chose to retain these words as 'neologisms' and see how they developed.

4.5.1 'Pre-Screening' and 'Pre-Exclusion' of Search Results

Since a number of unsuitable results inevitably almost always have to be removed from a corpus or database, I devised a system for 'pre-screening' search results in order to 'pre-exclude' webpages without needing to individually assess each one. This allowed for the creation of a much more nuanced, targeted database by avoiding the collection of non-relevant search results. It also precluded the need for time- and resource-consuming post-processing procedures.

This new process required that SRPs be manually read starting with the final page and working forwards towards the first (since it was found that duplicated and problem entries were generally located at the end of the list of search results). This enabled identification and 'pre-exclusion' of the following:

Duplicated articles: this was to deal with the fact that the same extract of text from the article could be used repeatedly, sometimes with slightly different article dates, sometimes with no differences at all (see Figure 4.14).

Figure 4.14: Duplicate entries on search results page for 'frenemy' article in the *Express*

These duplicates occurred because articles are republished every time a new reader comment is added. This was sometimes demonstrated by a number appearing in place of the usual date (top left, beneath the hyperlink) and sometimes by a changing number at the end of, or occasionally in the middle of, the URL. As the example above shows, sometimes there were hundreds of duplicates of the same article, in this case 687. In each instance, I checked random samples of the duplicated webpages to ensure that they were indeed duplicates, and the oldest entry on the search results page was selected, to ensure that I had included the earliest use of the word in my database. The duplicated articles were then 'pre-excluded' by running the GAS search again, but including a section of the repeating text from the SRP extract in the 'none of these words' field of the Advanced Search form. For instance in this case, 'Simon Cowell's my deadly frenemy' was used.

Neologism uses in links to other articles rather than in the article itself: this was to deal with the fact that the same extract would appear repeatedly beneath URLs for different articles, and when sample articles were checked, this was found to be a link to an article which had already been collected. The page containing the link did not contain the search word anywhere in the article itself. Since links to other articles did not qualify as 'journalistic writing' (see 3.5), these were 'pre-excluded' as above.

False positives: this was to deal with the fact that extracts of text showing a false positive would appear under article URLs on SRPs, for example 'washing machine' appeared repeatedly as a false positive for 'greenwashing'[91]. Terms which accurately

[91] It is interesting to note that this no longer occurs, suggesting that changes have been made to either the coding of the newspapers or the algorithms used in Google searches.

match the search word take preference over any similar word or phrase, therefore the appearance of false positives meant that there were no actual matches for the neologism in that article. 'Pre-exclusion' simply involved re-running the search with the false positive term excluded.

Archived articles: this was to deal with the fact that articles would effectively appear twice on the SRP list because some of the newspapers ran weekly round-up or archive pages. As these archives/round-ups also included the neologism, they triggered an additional search result. In *The Guardian*, these were easily identifiable by the URL that appeared on the SRP, appearing with a plus sign in the truncated URL below the hyperlink and sometimes in the SRP article title as well, as shown in Figure 4.15.

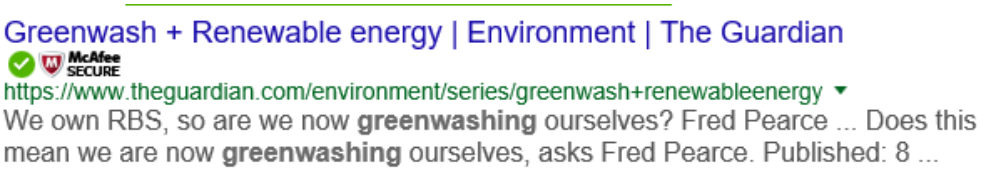


Figure 4.15: Archived article in *The Guardian* is indicated by the plus sign in both the article title and the URL beneath

'Pre-excluding' these articles in *The Guardian* simply required excluding the plus sign when the search was re-run. In the *Independent*, the word 'archive' appeared in the entry on the SRP and so 'pre-exclusion' was done in the same way[92]. Archived articles did not appear in the other two newspapers, meaning no action was required (although the *Mail* did return internal search results for some words, as discussed in 4.5.2).

Articles undated/published outside the required date range of January 2000 to August 2014: This was to deal with the fact that some search results contained dates in the extract below the hyperlink on the SRP, indicating that the article did not qualify for inclusion, and some search results contained no date, meaning that the article itself was undated (so that there was no way of knowing whether it fell into the required date range or not). These articles could not be 'pre-excluded' from the search, since GAS has

[92] This is another element of functionality which has since changed: archived articles in the *Independent* are no longer returned by a standard Google Advanced search

no ‘date range’ field, however such results were simply not collected in the next phase of the process.

Advertisements: this was to deal with the fact that search engines have traditionally included adverts relating to the user’s previous browsing history and their location (Fletcher 2013: 3), rather than relating to the search word. At the time of data collection, this meant that the adverts appearing on my SRPs had nothing to do with the neologism in question. As Figure 4.16 (captured in September 2016, in response to a GAS query for ‘cyberbullying’) shows, this has since changed, and adverts now do relate to the search word.

Figure 4.16: Advertisements carried a small ‘Ad’ logo in front of the hyperlink

Advertisements were identified through the white-on-green ‘Ad’ logo and their positioning at the top of the first page of results. These adverts still carry the same logo, but they now appear at the bottom of the first few pages of search results. It was not possible to ‘pre-exclude’ these advertisements, so they were simply not collected during the next phase of the project.

Proper Nouns: this was to deal with the fact that in some cases, neologisms were used as proper nouns, for example a race horse called ‘Rewilding’. Most proper nouns had been excluded before the final neologism list was completed; however those instances where a standard noun becomes a proper noun could not be planned for during the original selection process. The use of neologisms as proper nouns was clear from the text extract below the hyperlink, and hence those search results were not collected (since they could not be ‘pre-excluded’ without accidentally excluding required neologism uses).

Files which could not be opened: this was to deal with the fact that search results occasionally returned compressed files, requiring software such as WinZip[93] or

[93] See http://www.winzip.com/win/en/index.htm

WinRAR[94] to open them, or formatting and mark-up files, usually with a file extension like .xml.gz (see for example Figure 4.17), or 'Robots.txt' files.

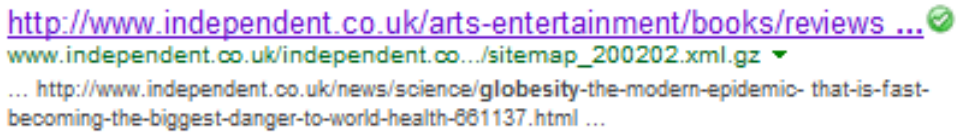


Figure 4.17: An extract from the *Independent* search results page for 'globesity', with '.xml.gz' file extension, indicating that this file either cannot be opened, or is a duplicate that is not required anyway

Most of these occurred in the *Independent* and were unlikely to be intended to be accessed by users. They probably only appeared on the search results pages due to a coding error. The abbreviation .xml is short for 'extensible markup language', and indicates that the file carries information only really of use to those designing, updating or managing the website (Fisher 2014). Based upon personal experience maintaining commercial websites, .xml files tend to include instructions to the rest of the website on how different components should appear, for example the font, size and styling of headlines, subheadings and bodycopy. Robots.txt files contain either duplicates of articles already collected, or (the majority) 'printer versions' of previously collected articles. The fact that they are picked up by GAS is believed to be a result of a coding error within the newspaper itself. (This view was corroborated by the fact that following the *Independent's* branding relaunch in 2015, these files no longer appeared in searches for the same neologisms.) It was not possible to 'pre-exclude' these articles through use of the GAS search form, and they were therefore simply not collected during the next phase of the project.

Following 'pre-screening' and 'pre-exclusion' (which often involved the inputting of several terms for inclusion in the 'none of these words' field of the GAS query form), new Search Results Pages were produced, ready for the final stage in the data collection process.

[94] See http://www.rarlab.com/

A slightly more complicated pre-screening process was required to exclude blogs and the few Reader Comments which had slipped through the process outlined in 4.2.3; these were generally neologism uses appearing in the first one or two Reader Comments, as they did still generate text extracts in search results. Most of the former had been dealt with (see 4.2.4 – 4.2.4.1), however a few slipped through the new procedures, for example as a consequence of 'minute-by-minute' reporting of sports events[95]. These were dealt with during the next stage of the data collection process (see 4.5.2 – 4.5.3).

The 'pre-screening' and 'pre-exclusion' processes discussed above demonstrate one way in which manual methods of data collection were found to be more appropriate to the development of context-rich genre-specific text collections. By avoiding the collection of non-relevant search results, and preventing the need for time- and resource-consuming post-processing procedures, they allowed for the creation of a much more targeted, and potentially larger list of qualifying search results to be collected.

4.5.2 Advance Exploration of Websites

Just as the 'pre-screening' discussed in 4.5.1 enabled me to 'pre-exclude' hundreds of search results without needing to open each file and individually assess each one, 'advance exploration' of webpages accessed from the Search Results Pages operated in a similar way. In this case, each file was opened, and certain characteristics were checked (and value judgements applied) before data harvesting began. If these characteristics met the criteria outlined below, the page was excluded without further analysis, and the files were not harvested for the database.

The 'advance exploration' criteria were as follows.

Broken links: as with any work on the web, a number of broken links were discovered. These usually carried a '404 server error' indicating that the link was broken. These URLs were checked three times over a period of several months and if the link remained broken they were excluded from the study.

[95] See for example http://www.theguardian.com/sport/2012/jun/08/england-west-indies-live-obo

Copyright/licence expired: some of the articles had appeared in the newspapers as a result of a licence being granted by another publication/author. Such licences are for a limited period only, and once they have expired, the article is removed from the website. It is usually replaced by a page explaining why the content is missing.

Internal search results: Google Advanced Search (GAS) regularly returned results from the *Express* which were themselves just Search Results Pages (SRPs) generated by the newspaper's own internal search engine. These would always include articles which had already been collected (hence representing duplicates) but bizarrely were based on search words completely unconnected to the neologism query which had generated them. Thus, for example, a GAS search for 'earworm' generated an *Express* SRP for 'Charlotte Heathcote', with an article (which had already been collected) reviewing a new CD appearing in the results list (see Figure 4.18).

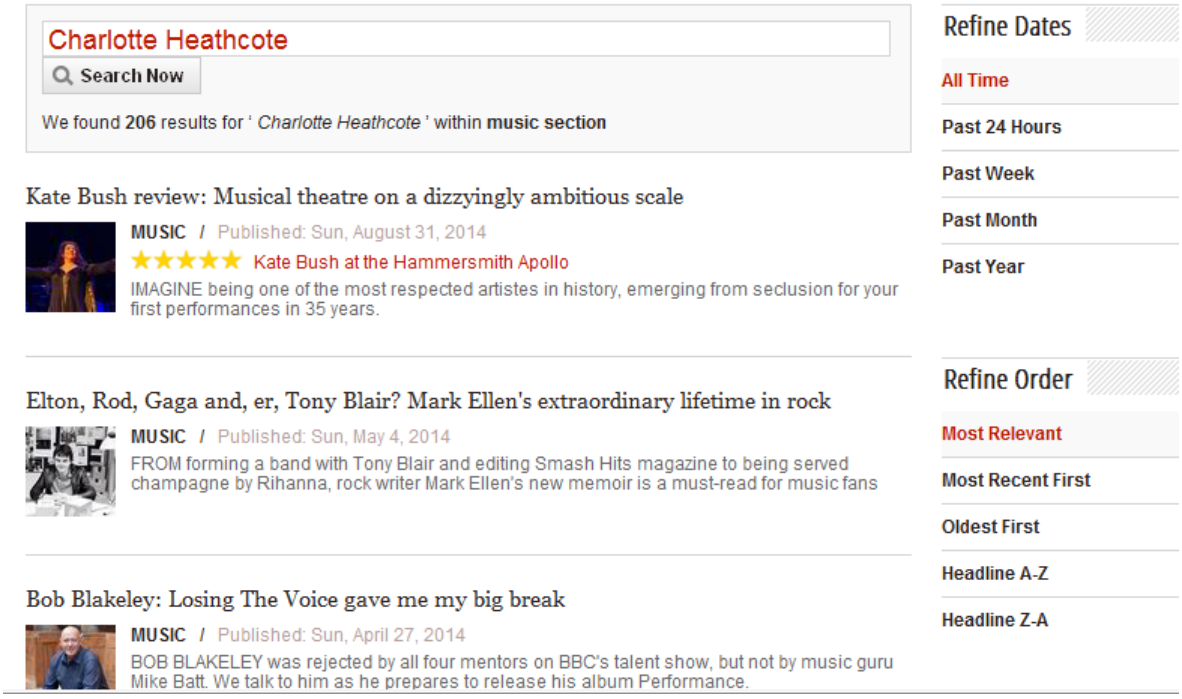


Figure 4.18: Internal search results from the *Express*, in response to a GAS search for 'earworm'

These strange search results proved intriguing, since there seemed no apparent connection between the search word used in GAS and the one appearing on the *Express* SRP. On discussing the issue with Coventry University's IT Services

department, it was determined that this was likely due to poor coding within the newspaper (Personal Communication, Coventry University IT Services 2014).

Neologism Articles: since the purpose of the study was to examine neologisms in use, articles about new words were considered individually, and interpretative criteria applied to determine whether or not they should be included. Articles were excluded where they were:

- Based entirely on a press release or other publicity material from a dictionary using new/unusual words to promote their products
- Words appearing in a list rather than as part of a full sentence[96].

Articles were included where the newspaper had used publicity material as a jumping off point from which to create its own article on language use, carrying out additional research on the use of the new words, or where the article was unconnected to any individual dictionary publisher, for example an article in the *Mail* offering advice to local government on not using 'goobledegook' like 'wellderly'[97].

Online newspaper versions: the purpose of this research was to examine online versions of print newspapers, however the *Mail* and *Express* carry several sections that only appear in the online version of the paper, for example the former's '*Mail Online'*[98] and articles in the latter marked 'for express.co.uk'[99]. As these articles had not appeared in the print versions of the newspapers, they were excluded from the study.

'Paid-for' articles: occasionally *The Guardian* prints articles paid for by outside organisations, which carry a banner or logo stating that they are 'paid for by ...'. These are known in the trade as 'advertorials' (a blend of 'advertisement' and 'editorial'),

[96] See for example https://www.theguardian.com/world/2009/jul/09/merriam-webster-dictionary-new-words
[97] http://www.dailymail.co.uk/news/article-1257097/Local-government-gobbledegook-phrases-banned-LGA.html
[98] See for example http://www.dailymail.co.uk/money/investing/article-2185618/JIM-MELLON-INTERVIEW-Britains-answer-Warren-Buffett-cracks-wealth-code-investing-book.html
[99] See for example http://www.express.co.uk/news/uk/326566/Italy-set-to-follow-Spain-as-both-countries-struggle-to-control-debt-and-borrowing-costs

since they carry the appearance of an editorial article, but the space attracts a fee. Since buyers of newspaper space are free to write as they wish (within limits), these were excluded from the study since they were not necessarily governed by the newspaper's normal style rules, and are not written by professional journalists (see 3.5.2).

Photos/videos: the database created here contains written English only so photos were excluded unless there was a caption containing the neologism. Videos were manually checked to ensure there was no textual component which could have contributed to the database (for example subtitles), however none were found.

Press agencies: many articles carry bylines stating that the article has been written by a news agency (for example Reuters[100], Press Association[101] or Associated Press[102]. These agencies provide news to outlets worldwide, and their stories may be printed 'as is', or expanded upon by the newspaper's own journalists (in which case they would generally carry the journalist's name). As one of the purposes of this project was to compare different newspaper's use and treatment of neologisms, press agency articles were excluded, since they were likely to be common to all four newspapers.

Question & Answer Features: these were most commonly found in *The Guardian*, and took several different forms. Those where the answer was written by the interviewer based on comments made by the interviewee were acceptable to the project, since the interviewer was a professional journalist. However those in which the answers were the actual words of the interviewee, indicated by short, colloquial sentences, (for example '*This Much I Know – Joe Calzaghe*'[103]) were excluded since they were not written by professionals.

Reader Letters: just as reader comments were not considered relevant to this study because they were not produced by professional writers adhering to language rules set by the newspapers' editors/publishers, so reader letters were similarly excluded.

[100] See http://uk.reuters.com/
[101] https://www.pressassociation.com/
[102] http://www.ap.org/
[103] http://www.theguardian.com/lifeandstyle/2010/aug/08/joe-calzaghe-boxer-this-much-know

These occurred most frequently in the *Independent* newspaper, in its *Voices* section.[104]

Speeches reproduced verbatim: occasionally political or business speeches were printed (usually in *The Guardian* or the *Independent*) in full[105]. Again, since these did not meet the required criteria of journalistic writing, they were excluded from the study.

Blogs: due to the potential for confusion previously identified over whether a text in *The Guardian* was an article or a blog (4.2.4) blogs were added to this list, albeit with a slightly altered method for exclusion (see below).

All search results matching these criteria were excluded from the study and the URLs leading to the articles were therefore not collected.

Also not collected were the search results identified during 'pre-screening' which, for various reasons, could not be 'pre-excluded' (see 4.5.1):

- Undated articles
- Advertisements
- Files which could not be opened – those with the file name Robots.txt or the file extension .xml.gz
- Proper nouns such as the race horse named Rewilding
- Blogs

[104] See for example http://www.independent.co.uk/voices/letters/letters-turning-a-blind-eye-to-the-butchery-in-gaza-9624303.html
[105] See for example http://www.theguardian.com/science/2011/apr/06/templeton-prize-2011-martin-rees-speech

4.5.3 Methodology for the 'Pre-Exclusion' of Blogs

In this section I outline the slightly revised methods established for preventing collection of blogs, particularly in *The Guardian*. This was done by amending the search parameters used within Google Advanced Search (GAS), as per the 'pre-screening' and 'pre-exclusion' processes discussed in 4.5.1.

A GAS search for the word 'blog' used in the 'none of these words' field would result in all articles containing the word 'blog' being excluded, but did not permit the exclusion of the various 'Blogs' mentioned in 4.2.4.1 without also losing every article that mentioned the word 'blog' in its text. Clearly this was unacceptable.

Alternatively, the names of the blogs which had been identified (such as 'theatre blog' and 'media monkey blog') could be used in the exclusion field. However this would result in repeated failed searches due to the number of blogs that needed to be excluded and the character limit placed on this field in the search parameters form. Thus the only way to exclude all *Guardian* blogs whilst retaining relevant *Guardian* articles was to consider each neologism individually, identify which *Guardian* blogs the term was most likely to appear in, and exclude these specific ones from the search parameters. Thus for 'greenwashing', 'sustainability' blogs were excluded, and for 'hyperlocal' media blogs. For any blogs which made it through this filtering process, for example where there were still too many potential blogs to fit in the search parameters' exclusion field, advance exploration was applied to the webpages, using the blog criteria established in 4.2.4.1.

While it is not possible to tell how many blogs posts were excluded following the 'pre-screening' of search results (that is, named blogs included in the 'none of these words' search field), it is known that 711 blogs were excluded through the advance exploration of websites.

It is likely that hundreds more were excluded through standard 'pre-screening', since it is known that 297 were excluded in this way for 'cyberbullying', 225 for 'hyperlocal' and 327 for 'sovereign debt'. This totals 849 (1560 when we include the 711 mentioned above).

4.5.4 Media Tracking – Harvesting URLs and Collecting Data

In this section I outline the process undertaken to collect the newspaper articles which had passed 'pre-screening' and 'advance exploration' of websites.

Having completed all stages of preparation of the search results ready for manual collection of articles, it was found that, through 'pre-screening', 'pre-exclusion' and 'advance exploration' of websites the number of qualifying articles was reduced from the 9,200 produced by the initial Google Advanced Search to just under 4,000. Over 5,000 non-relevant results were therefore removed from the list, saving enormous time and resources in the final data collection process.

In summary, the data collection procedure was as follows:

- Articles were collected by newspaper, as per the Google Advanced Search process

- Search result hyperlinks were clicked, leading through to the target article

- URLs for each article were first copied and then pasted into a Microsoft Word file. All subsequent information was added to this file. The Word file was later converted into a Microsoft Excel spreadsheet, however at this stage Word was the more appropriate program, since a) it was less cumbersome for the collecting of data than Excel and b) it was easier to prepare the information for upload into a Corpus Query program from Word than from Excel (see 4.6.2). The information in Word was, however, formatted in such a way that it could easily be converted into a table, which would then convert seamlessly into an Excel worksheet.

- Year of publication was added to the information in Word. Neologisms were tracked in newspapers dating back to 2000, in order to correspond with the earliest of the Dictionary Date of Entry Batches. However the earliest use of the word in each newspaper was also collected, if this was pre-2000, allowing for an extended picture of the life of neologisms to be presented (see 5.2).

- Full date of publication was noted down, so that month-by-month information was available.

- Type of article was noted, for example 'news', or 'lifestyle'.

- Number of neologisms on the page was counted, using the web browser's 'find on this page' function, and those appearing outside of the main article (for example in links or in Reader's Comments) were ignored. (Although articles containing neologism use only in Reader Comments had been excluded from the study, those with instances both in the main text and in the comments section were not. Therefore the Reader Comments had to be removed at this stage instead.) A note was however made each time there were extraneous neologism instances, to avoid potential future confusion, should the 'find on this page' function be used again on the page, and the numbers were not to tally.

- Neologism instances appearing in headlines were marked as a component of the total number of uses.

- Spelling variants (see 4.3.2.1) were noted (each having been counted in the same way as all other neologisms), whether they were in a) articles using solely one variant or b) articles using one or more variant spellings. Where the latter was the case, a note was made to that effect.

- Additional comments about the article were marked, for example the fact that the neologism appeared in a 'pull-out' box which provides additional information on the topic.

4.6 Creating the *NTON* Database

Having completed this stage of data collection, the next step was to tidy all of the entries and prepare for transfer into the software from which the database would be analysed. In this case, there would be two versions of the database, one in Excel, where

all the frequency analysis would be conducted, and one in a corpus query program, where corpus analyses would be done.

4.6.1 Neologism Tracking in Online Newspapers: Excel

The decision was made to form two versions of the database because it was discovered that the Sketch Engine article list contained 81 fewer articles than the Excel version of the database. This was believed to be due to coding or output errors within Sketch Engine, since the number of articles and neologisms had been repeatedly verified in Excel. Such verification is more complex in Sketch Engine; a number of articles were identified as missing from the list produced by the program, yet their contents were found in the database by using the concordancing software to find lines of text known to be in the 'missing' articles. Thus it is likely that the error lay in the article list output and not in the actual database. However since the pattern of articles per newspaper was the same regardless of which program was used, and the conclusions drawn based on this data are unaffected by program choice, this discrepancy is not considered problematic. Indeed discrepancies between differing corpus query programs are common, and it is widely accepted that the researcher must choose the program which offers the figures believed to be most accurate.

For the Excel version of the database, the contents of the Word file were converted into a table and the table was then imported into Excel. Each neologism was saved into a different file, so that a variety of searches could be run on different worksheets, for example sorting by date, by newspaper, by article type and by number of neologisms. One file containing all component neologisms was created for each of the datasets organised by date: Dictionary Date of Entry Batch 1+2 (DDEB1+2) and DDEB3.

Before the Sketch Engine version of the database was compiled, the issue of articles containing spelling variants was addressed. As discussed in 4.3.2.1, these had been harvested in the same way as every other neologism, however because the words with spelling variants were harvested from two different sets of search results this meant that where more than one variant was used in the same article, that article appeared twice in the database – once for the compound version of the word and once for the

hyphenate/two-word term. It was therefore necessary to ensure that whilst all neologism instances were maintained, the database only contained one entry for the article. This was done using the 'conditional formatting' feature in Excel, which allows the user to highlight duplicated text, in this case URLs. Having identified all of the duplicate pairs, the database was manually reviewed and one of each pair was marked as '0' in the 'article' column, whilst the 'neologism instances' column remained unchanged (since this showed the appearances of the word for that spelling variant). This was done both on the individual Excel files and on the combined DDEB1+2 and DDEB3 files. In addition, the database was further tidied and refined, correcting any other minor errors. The Excel DDEB1+2, and DDEB3 spreadsheets were then converted for upload via WebBootCaT, as discussed below.

4.6.2 Sketch Engine Database: Challenges and Solutions of Uploading URLs through WebBootCaT

Sketch Engine (SkE) was used for the analyses due to its reputation as 'a leading tool for lexicography and other corpus work' (Kilgarriff and Kosem 2012: 32). Sample data runs were prepared for upload using WebBootCaT (WBC). This involved similar 'post-processing' to that outlined by Fletcher (2013: 5). Files containing 10 neologism article URLs were stripped of all formatting and converted into plain text (.txt) files. They were then loaded into WBC, following the instructions on bulk uploads on the Sketch Engine FAQs page[106].

A number of initial difficulties were encountered, particularly with file sizes and with the *Independent* newspaper timing out before WBC could download its files (error message 'failed to retrieve').

Staff at SkE proved very helpful in addressing these remaining problems, and in directing me to the 'autolog' files, which contained error messages explaining the remaining failures. Vit Suchomel on the Technical Support team supplied all of the necessary answers (as well as advising me that my entire corpus datasets could be

[106] https://www.sketchengine.co.uk/ske-faq-frequently-asked-questions/. The Sketch Engine site has been updated since I uploaded the NTON database, and many pages, including this one, have changed significantly since then.

uploaded at once; not 100-at-a-time as the guidance notes had suggested (Personal Communication, July 2015). Having adopted Suchomel's suggested changes, a test run was carried out for Dictionary Date of Entry Batch 1 and 2 (DDEB1+2), comprising 100 files. This resulted in just 13 failed files – four 'unable to retrieve' and nine 'duplicates'.

A single upload into WebBootCaT was then carried out for each of the Dictionary Date of Entry Batches, including all of the URLs for each dataset. This resulted in 387 failed files, 162 from DDEB1+2 and 225 from DDEB3. Working through the autologs allowed for identification of the failed files; duplicates were discarded, to avoid any skewing of results, and 'unable to retrieve' (49 in DDEB1+2 and 201 in DDEB3, all due to the *Independent* timing out) were put into new groups and retried repeatedly over the next three days, each time a few more being accepted. Finally, a handful were left from DDEB3 which Sketch Engine was simply unable to upload; these were manually entered into the corpus by opening the page, copying the article text and uploading it through the main upload procedure, rather than through WebBootCaT.

4.7 Gathering and Analysing Dictionary Comparison Data

In this section I briefly outline how entries from the five dictionaries under study here were collected and organised so that a comparative analysis could be conducted on their various components.

Throughout, the focus lay upon answering Research Question 1:

> *What can be learnt from this study about* Wiktionary's *responsiveness to neologisms and the level of detail and quality of definitions in its new word entries, when compared with expert-produced dictionaries?*

This supported Objective 1: comparing degrees of comprehensiveness between expert-produced dictionaries and *Wiktionary*.

Each of the 26 neologisms known to have already entered a dictionary – and thus being members of either DDEB2 or DDEB3 – was checked against each of the five

dictionaries, and a detailed Excel spreadsheet was drawn up for each dictionary, showing the standardised and non-standardised components explained in 3.4.3 and listed in Table 4.10.

Dictionary Component	**Description**
Standard Dictionary Components	
Headword (lemma)	Indicator of how a word is written
Lexical unit	Subdivisions of headwords (senses)
Menu	List of lexical units
Definition	Explanation of the meaning of a headword
Pronunciation	Guidance on how a word should be pronounced
Etymology	Origin of a word
Spelling variant	Variations in spelling of a word
Word class	Part of speech
Grammar label	Indicator of grammatical information on the headword
Register/style/attitude labels	Indicators of the type of word
Domain label	Marker of the field to which the headword applies
Region label	Indicator of where the word is generally used
Example	Text elucidating the meaning, illustrating use or attesting to the presence of a headword in the language
Usage note	Notes giving additional information on using a headword
Cross-reference	Indicator that more information is available elsewhere
Run-on	Indicators of words derived from the headword
Non-Standard Dictionary Components (mainly used in *Wiktionary*)	
Inclusion date	Indicator of when the word first entered the dictionary (also provided for some words in *OED*)
Revision History	Save-by-save record of every change ever made to an entry
Discussion Forum	Online spaces for discussion of dictionary entries, specifically Talk pages and the Tea Room
Audio File	Sound file added to help with pronunciation (now found in many electronic expert-produced dictionaries)
Translation	Headwords provided in multiple languages (*Wiktionary* only)
Derivative	Marker that the neologism under study derives from another headword, for example in *OED* 'cyberbullying' is a derivative of 'cyber'
Related term	Indicator that a word is linked to the headword, although it is not an actual run-on. In standardised dictionaries this might appear as a Usage Note, however it is treated separately here as *Wiktionary* treats it as a separate element
Synonym	Word which means the same as the headword. Also often included in Usage Notes, but separated here for the same reason as related terms
Contents navigation panel	Panel of hyperlinks to help users move around longer entries in *Wiktionary*

Table 4.10: List of standardised and non-standardised dictionary components

For each dictionary, where a neologism was found to include one of the components in Table 4.10 this was marked into a binary system of 1s and 0s on the spreadsheet, indicating the presence or absence of the component. These were tallied for each neologism, for each dictionary and for each component, in order to gain insight into the degrees of comprehensiveness of each publication.

These frequencies were analysed both individually and in concert, for example the *Oxford English Dictionary* and *Merriam-Webster* dictionary were analysed together to gain a picture of the position of 'corpus-informed' dictionaries (see 3.4.1). The *Oxford Dictionary of English* and *Oxford Dictionaries* online were also analysed together, since they are both 'corpus-based', and *Wiktionary* was analysed alone, as the only collaborative dictionary in the study.

One of the key elements of these comparisons was that of the dictionary definitions. These underwent subjective qualitative comparison, assessing the defining style used (see 3.4.3) and the degree to which the definitions matched one another in meaning.

The information from these dictionaries was also compared with that in the *Oxford English Corpus* (*OEC*) since this would shed further light on the responsiveness – or otherwise – of these publications to new words.

All of this allowed conclusions to be drawn about the speed of response of the various dictionaries, as well as the efficacy of corpora in providing information for dictionary entries.

4.8 Conclusion

In this chapter I outlined the selection of neologisms for use in this study and the many processes, including the Research Randomizer undertaken to arrive at a final list of 34 new words. I discussed the initial culling of the *NeoCrawler* list of neologisms, to exclude words such as proper nouns and those which in reality were more likely spelling errors than actual new terms. I further discussed testing of these words against newspapers, and the issues this raised in terms of false positive results and

inconsistencies in the information available through internal search engines. I explained the subsequent decision to utilise external search engines instead, since these would enable the same criteria to be applied to every neologism in every newspaper. I also explored the challenges presented by new international legislation that can result in web-based articles disappearing from sight without warning, and those surrounding the potential confusion caused by neologisms appearing in newspaper blogs rather than articles written by professional journalists.

I moved on to discuss automated and semi-automated systems of corpus data collection, a process undertaken in order to demonstrate how the manual methods devised in this study were in fact more suitable to the task. These I described in detail, introducing the concepts of 'pre-screening' search results pages to remove unsuitable candidate texts before ever downloading their contents, and 'advance exploration' of websites, designed to identify and utilise contextual information such as, in this case, date. This kind of 'pre-exclusion' also helps a corpus data collection project to be more closely targeted, since it allows the researcher to see, for example, that the article is written by a press agency or is the transcript of a speech (and hence is likely to be used in the same format in competing newspapers).

I then described the process of collecting the corpus texts which *were* deemed suitable for inclusion, and the creation of the database itself, including the decision to create both an Excel and a Sketch Engine version of the database, so that a variety of corpus analyses could be carried out.

Chapter 5 Findings and Discussion

5.1 Introduction to Findings and Summary of Findings and Discussion

This chapter presents and discusses findings obtained both through the analysis of data gathered using the new methodology outlined in Chapters 3 and 4, and through examination of a range of five dictionaries of different types and formats. These findings seek to answer the Research Questions established in 3.9, whilst also addressing the objectives of the study.

As established in 1.1, the objectives of this study were initially two-fold:

1. To compare degrees of comprehensiveness in the entries provided for new words in expert-produced dictionaries with those in collaborative dictionary *Wiktionary*

2. To track neologism appearances in UK news media in order to compare usage and behaviour in different newspapers, at different stages in the neologic life-cycle

As noted above, achieving Objective 2 involved the design, creation and implementation of a new method of data collection, which was aimed at creating context-rich genre-specific corpora. Since this was an exploratory study, with the use of the new methodology serving as a pilot for its future use and development by other researchers, a third objective was added to the original two:

3. To consider whether neologism use and behaviour in the media can be best explored through the use of new manual or existing automated corpus data collection techniques.

5.2 An Overview of Neologism Use: Datasets, Dictionary Entries and Media Appearances

In this section, I provide an overview of the factors influencing the findings to be presented in this chapter, plus a number of findings which apply across all elements of the study. These factors relate to both the dictionary comparison and media tracking elements of this study.

As discussed in 3.4, the dictionaries used to address Objective 1 were:

- *Oxford English Dictionary (OED)*
- *Oxford Dictionary of English (ODE)*
- *Oxford Dictionaries* online (*ODO*)
- *Merriam-Webster* (*MW*)
- *Wiktionary* (*W*)

Analysis of these dictionaries was mediated through their relationship to corpora, each dictionary being categorised according to whether it was 'corpus-based, 'corpus-informed' or 'collaborative'. Taking account of these groupings, the dictionary list looked like this:

- Corpus-based dictionaries:
    - *Oxford Dictionaries* online (*ODO*) (2014)
    - *Oxford Dictionary of English* (*ODE*) (Printed book (2010))

- Corpus-informed dictionaries:
    - *Oxford English Dictionary* (*OED*) (online) (2014)

- *Merriam-Webster* (online) (2014)

- Collaborative (corpus-free) dictionary:
  - *Wiktionary* (online) (2014)

As discussed in 3.4.4, the neologisms were separated into three datasets, based upon when the neologisms within them first entered a dictionary and encompassing the entire neologic life-cycle. The first two of these were in most instances combined (appearing as DDEB1+2) since they covered the same timeframe. When it was necessary to examine the three individually (for example when conducting comparisons of dictionary entries or during media tracking) the first dataset was either excluded, or treated as a separate category. Thus the three datasets (which are laid out in full in 3.4.4) comprised the following:

- Dictionary Date of Entry Batch 1 (DDEB1) (not yet appearing in a dictionary)
- Dictionary Date of Entry Batch 2 (DDEB2) (most recent neologisms)
- Dictionary Date of Entry Batch 3 (DDEB3) (most well-established neologisms)

These datasets would allow for the comparisons of degrees of comprehensiveness in dictionary entries provided for new words (required for Objective 1: see 5.1) looking both at individual date-groups (for example DDEB2), and across them (for example, whether neologism entries ever change over time). The datasets would be even more important for the media tracking project, since Objective 2 specifically requires that usage be examined at different stages (DDEBs) in the neologic life-cycle.

Thirty four neologisms were selected for use in this study (see 4.2 and its subsections); eight reside in DDEB1, 11 in DDEB2 and 15 in DDEB3.

The use of these new words was tracked in four UK national newspapers, in order to address Objectives 2 and 3 of this project. The newspapers were:

- *The Guardian*
- *Independent*
- *Mail*
- *Express*

Using the new methodology devised during this project, a 4.2million word database entitled *NTON* (*Neologism Tracking in Online Newspapers*) was created, comprising articles from these newspapers containing at least one instance of one of the 34 neologisms mentioned above.

The expectation was that words in DDEB3 would exhibit higher levels of media use, and more comprehensive dictionary entries than words in DDEB1+2 combined, simply by virtue of the fact that they dated back to the earliest points of the neologic life-cycle, having been in existence for as many as 14 more years.

As Table 5.1 indicates, DDEB3 entries carry, in total, three times as many components as the entries in DDEB2 (DDEB1 not yet having entered dictionaries, and therefore having no entries to compare). However (working in relation to date), DDEB1+2 (the most recent stage in the neologic life-cycle) holds more neologisms than DDEB3 (the earliest) (19 versus 15) and also more articles (1,947 versus 1,926) and more neologism uses (3,049 versus 2,891) than DDEB3.

	DDEB1+2 and DDEB3	DDEB1+2 Neologisms not yet appearing in *Wiktionary* / any expert-produced dictionary as at 31 August 2014 AND Neologisms first entering *Wiktionary* / an expert-produced dictionary between September 2008 and August 2014			DDEB3 Neologisms appearing in *Wiktionary* / an expert-produced dictionary between January 2000 and August 2008
	DDEB1+2 and DDEB3 Total	DDEB1 + DDEB2 Total	DDEB1 Neologisms not yet appearing in *Wiktionary* / any expert-produced dictionary as at 31 August 2014	DDEB2 Neologisms first entering *Wiktionary* / an expert-produced dictionary September 2008-August 2014	DDEB3 Total
NTON Tokens	4,051,383	1,820,265	137,676	1,682,589	2,231,118
Number of neologisms	34	19	8	11	15
Number of Neologism uses	5940	3049	133	2916	2891
Number of Articles featuring neologisms	3873	1947	117	1830	1926
Number of Dictionary Entry Components	567	134	N/A	134	433

Table 5.1: Spread of neologisms across DDEB1, 2 and 3

Thus the only area in which reality holds true to expectations is in the number of dictionary components, where we see that there are significantly more components for the more established neologisms. This suggests that over time entries do expand and become more comprehensive.

Although the differences in the numbers of neologisms/articles are not substantial, it is at first glance surprising to see that the words which had been in dictionaries (and the neologic life-cycle) the longest (DDEB3) (and hence might be considered better established) occurred least frequently in the media. However it must be borne in mind that DDEB1+2 contains four more neologisms than DDEB3; three of the words in DDEB2 are responsible for 89.6% of the total number of neologism appearances. These neologisms are present in such high numbers due to external factors which will be discussed in 5.4.3; without these factors, the pattern of neologism and article numbers would fit with my original expectation: the best established category of

neologisms (DDEB3) would hold the highest number of neologism uses and articles containing neologisms.

The spread of neologisms across the 14 years of the study is illustrated in Tables 5.2 and 5.3, and represented graphically in Figures 5.1 and 5.2. The spread is presented in terms of when individual new words entered particular dictionaries, and the numbers of newspaper appearances in different years. As will be the case in most analyses, DDEB1 and 2 are presented together, although they are colour-coded to aid identification of the two categories. The graphics will provide a useful context for the findings presented throughout this chapter.

DDEB1	DDEB2																		
	2014	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004	2003	2002	2001	2000	Entered OED	Entered ODO	Entered MW	Predates NTON Corpus: earliest use in newspaper	Newspaper Name
Bankster	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0		After 2010			
Buzz marketing	0	0	1	0	2	0	2	2	0	1	4	1	0	0	0					
Cold peace	0	3	4	8	0	0	0	0	3	3	0	0	1	0	0				26.11.93	*Independent*
Cyberbullying	168	370	211	155	55	86	71	66	13	1	0	0	0	0	0	After 2009	After 2010	Date unknown	06.03.99	*Guardian*
Cyberchondriac	0	0	0	0	1	7	2	1	0	0	1	0	0	0	0	After 2009	After 2010		28.06.98	*Independent*
Diabesity	1	1	9	0	0	0	0	0	0	0	0	0	0	0	0		After 2010			
Floordrobe	1	1	1	0	1	0	1	0	0	0	1	0	0	0	0					
Gendercide	1	1	2	3	8	0	0	1	3	0	0	0	0	0	0				01.12.96	*Independent*
Globesity	0	0	0	4	1	0	3	0	1	0	0	0	1	0	0					
Hyperlocal	14	51	102	42	52	24	5	2	0	0	0	0	0	0	0		After 2010			
Newer markets	0	3	4	4	2	2	3	0	1	1	0	1	1	0	0				25.09.94	*Independent*
Open education	0	2	1	3	0	1	0	0	0	0	0	0	0	1	0				14.11.96	*Independent*
Predatory lending	2	2	4	1	8	3	0	4	2	1	0	4	0	0	0					
Rewilding	10	41	7	7	5	4	4	3	2	9	0	0	0	0	0	After 2010				
Round pound	0	0	0	1	2	5	1	0	0	0	0	0	0	0	0				19.02.95	*Independent*
Sodcasting	0	0	1	1	2	0	0	0	0	0	0	0	0	0	0					
Sovereign Debt	38	67	275	601	225	29	1	0	0	0	7	0	0	1	0		After 2010		06.01.98	*Independent*
Superphone	5	12	4	0	6	0	0	2	1	0	1	0	0	0	0				25.06.98	*Independent*
Tablet computing	1	3	9	7	4	0	0	0	0	0	0	0	0	0	0					

Table 5.2: DDEB1+2 Most recent elements of neologic life-cycle of neologisms in newspapers (raw data)

Key:

Entered Wikitionary
Entered OED
Entered ODO
Entered MW

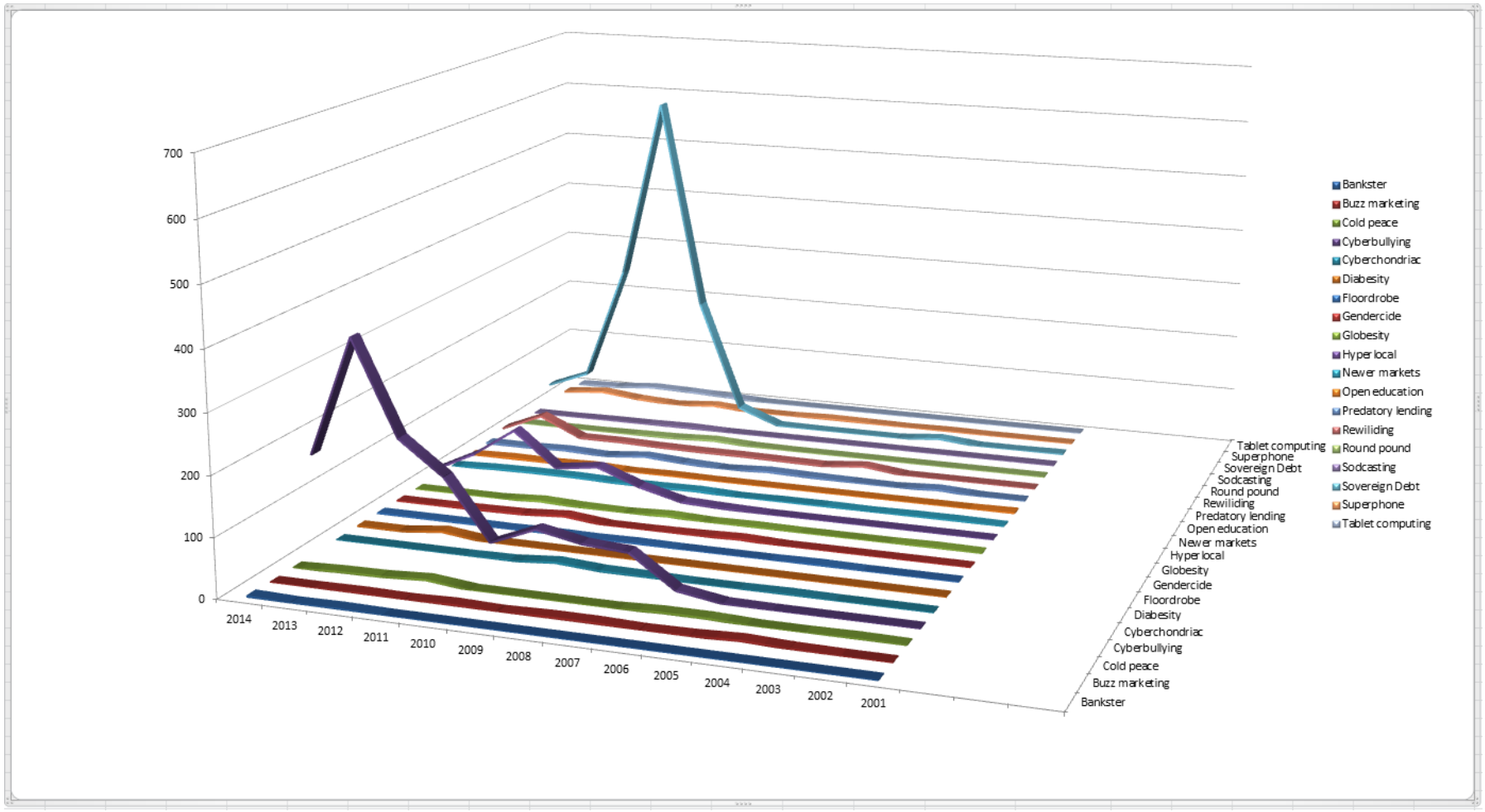


Figure 5.1: DDEB1+2 Most recent elements of neologic life-cycle of neologisms in newspapers (raw data)

	2014	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004	2003	2002	2001	2000	Entered OED	Entered ODE / ODO	Entered MW	NTON Corpus: earliest use in newspaper	Newspaper name
Acedia			1	1		2										1933	Included in1st Edition, 1998	Date unknown	16.02.97	*Independent*
BOGOF	10	16	25	14	23	34	11	2	1	1	1	2	1						13.03.94	*Independent*
Conurbation	10	35	43	28	21	30	44	23	15	17	23	12	11	6	9	1972	Included in1st Edition, 1998	Date unknown	13.08.92	*Independent*
Earworm	12	17	40	11	13	5	6	3	28	2	2				4			Date unknown	03.06.97	*Independent*
E-tailer	23	23	12	4	8	6	5	7	2	2				2	18			Date unknown		
E-waste	4	3	5	11	6	11	6	1					1		1			Date unknown		
Frenemy	12	3	12	12	2	3	2	1										Date unknown		
Greenwashing	8	18	10	7	24	16	15	7	4	4	4	1						Date unknown	17.08.94	*Independent*
Hubristic	29	61	57	64	39	46	38	22	26	17	14	11	9	4	4	1899	Included in1st Edition, 1998	Date unknown	01.02.90	*Guardian*
Promissory note	3	5	5	7	1		2	1		2	1	4	1		1	Date unknown	Included in1st Edition, 1998	Date unknown	25.03.93	*Independent*
Tenebrous	1	2	14	6	4	6	2	3	2	1	1	2		2	1	1911	Included in1st Edition, 1998	Date unknown	15.08.93	*Independent*
Upskill	1	6	6	7	5	5	6	1		2		1		1		1993				
Warrantless	5	37	16	6	3	8	4	3	5	1						1921			21.11.93	*Independent*
Waterboarding	74	143	185	219	246	152	154	75	24	8	1							Date unknown		
Wellderly		1	4	3	11															

Table 5.3: DDEB3 Oldest elements of neologic life-cycle of neologisms in newspapers (raw data)

Key:

Entered Wikitionary
Entered OED
Entered ODE/ODO
Entered MW
Updated OED

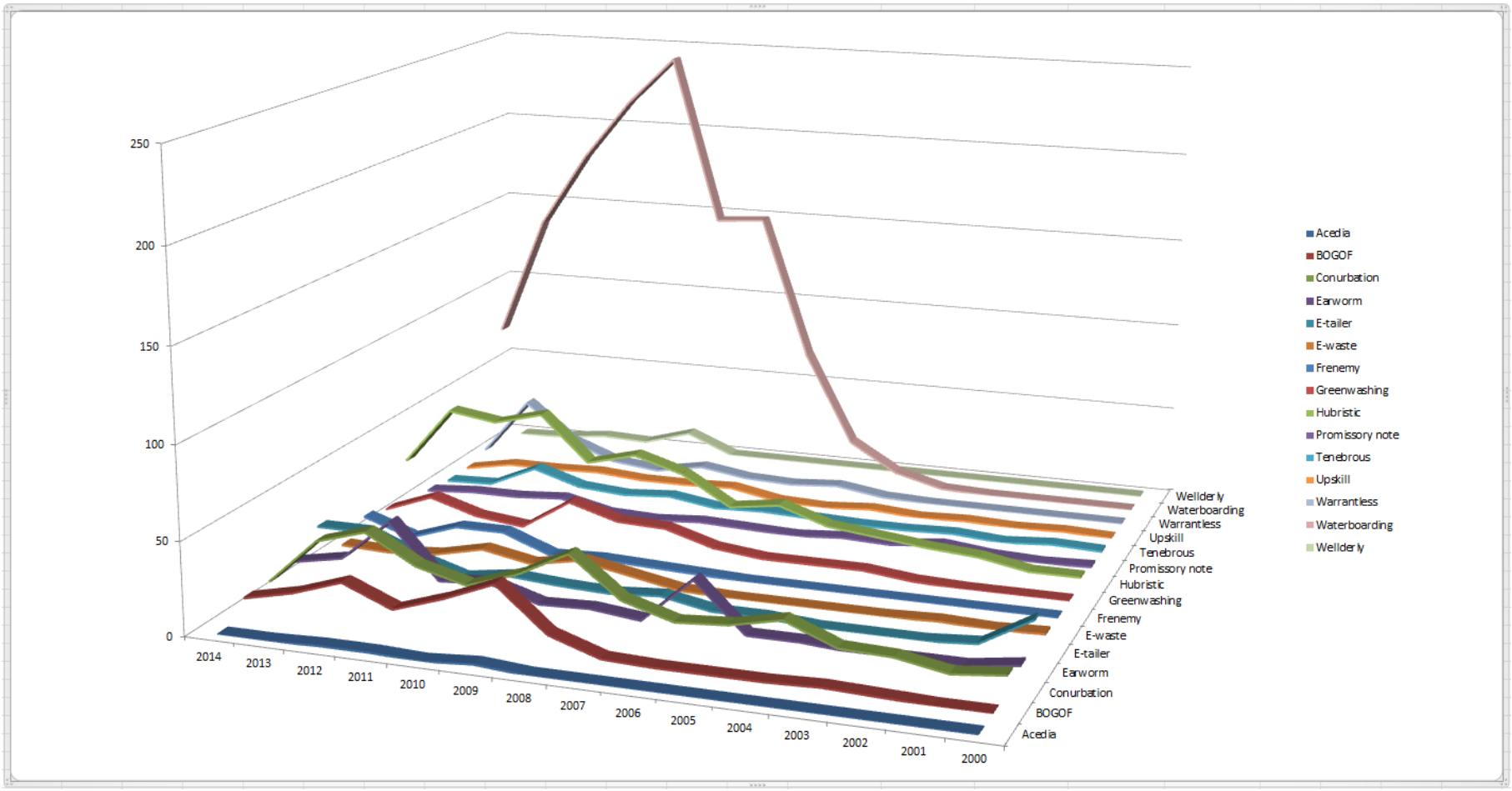


Figure 5.2: DDEB3 Oldest elements of neologic life-cycle of neologisms in newspapers (raw data)

In Table 5.2 we can see that the words in DDEB2 (those having entered a dictionary between 2008 and 2014) had all appeared in *Wiktionary*, but many of the words in DDEB3 had also already appeared in expert-produced dictionaries. Indeed, to be accepted into expert-produced dictionaries the requirements of the attestation process (see 3.4.2) would probably have resulted in them having a history of use in the media.

As far as it is possible to tell, given the limited and at times unreliable dating information available, the majority of these words have been present in expert-produced dictionaries for some years.

In Tables 5.2 and 5.3, where the colour coding is marked as 'date unknown', this indicates that the dictionary did not include date information. In DDEB3, where the *Oxford English Dictionary* date is shown as, for example 'after 2009' this is because the neologism is a derivative of another word, and the only information available is that the word from which it was derived entered in 2009. Words shown as entering the *Oxford Dictionary of English (ODE) / Oxford Dictionaries* online (*ODO*) in 2003 and 2010 entered the printed edition of the *ODE*. Since *ODO* provides no publication date, entry dates for *ODO* are unclear. In DDEB2, 'after 2010' indicates that the word entered *ODO* since the 2010 publication of *ODE*.

It is likely that some of the neologisms in DDEB3 which the *NeoCrawler* had identified as 'new' had actually been in use for many years. According to Tables 5.2 and 5.3, all but five of these neologisms showed a marked increase in usage in the years following entry into *Wiktionary*. This occurred in 2005-2008, around the same time as the *NeoCrawler* cut-off date, and suggests that although they may have experienced previous usage, they were on the cusp of entering a new phase in their own life-cycle.

5.3 Contrasting Representations of Neologisms: Lexicographical Perspectives

In this section, I present and discuss findings which address Objective 1 of this study – to compare the comprehensiveness of entries provided for new words in expert-produced dictionaries with those in collaborative dictionary *Wiktionary* – setting out differences in the ways in which different dictionaries and different dictionary types approach the representation of new words within their pages. These findings also allow me to draw conclusions in answer to Research Question 1:

> *What can be learnt from this study about* Wiktionary's *responsiveness to neologisms and the level of detail and quality of definitions in its new word entries, when compared with expert-produced dictionaries?*

Before beginning detailed presentation of my findings, I offer first a brief summary, which outlines what in fact can be learnt from this study about *Wiktionary*'s responsiveness, level of detail and quality. *Wiktionary* was found to be the most comprehensive of the five dictionaries under study here, in terms of the dictionary components contained within its entries (both number and quality), the speed and manner in which it can respond to new or changing neologisms, and the nature and quality of its definitions. *Wiktionary* achieves greater detail in its new word entries partly because it is not bound by the conventions of standardised dictionary structures. Thus it can include both standardised dictionary components such as pronunciation guidance, usage notes and word class markers, and non-standard ones such as revision histories, inclusion dates and examples drawn from sources that would not be permitted in traditional dictionaries. Also important is the way in which contributors build upon each other's work, pooling their resources and their knowledge. A key part of this is the discussion culture, which allows *Wiktionary* to respond much more quickly to new words than expert-produced dictionaries can do. The Revision History function means that *Wiktionary* is updated hundreds of times a day, and this also results in a greater responsiveness to neologisms, since changes to an entry that might take months to appear in an expert-produced dictionary can do so in *Wiktionary* in mere hours.

5.3.1 Neologism Inclusion in Dictionaries

Of the 34 neologisms appearing in dictionaries in this study, the largest number appear in *Wiktionary*, as shown in Table 5.4 and Figure 5.3, although *Wiktionary* ties with *Oxford English Dictionary* (*OED*) for DDEB3.

Dictionary	DDEB2	DDEB3
Oxford English Dictionary	3	12
Oxford Dictionary of English	0	14
Oxford Dictionaries online	6	13
Merriam-Webster	1	11
Wiktionary	9	14

Table 5.4: Number of neologisms appearing in each dictionary

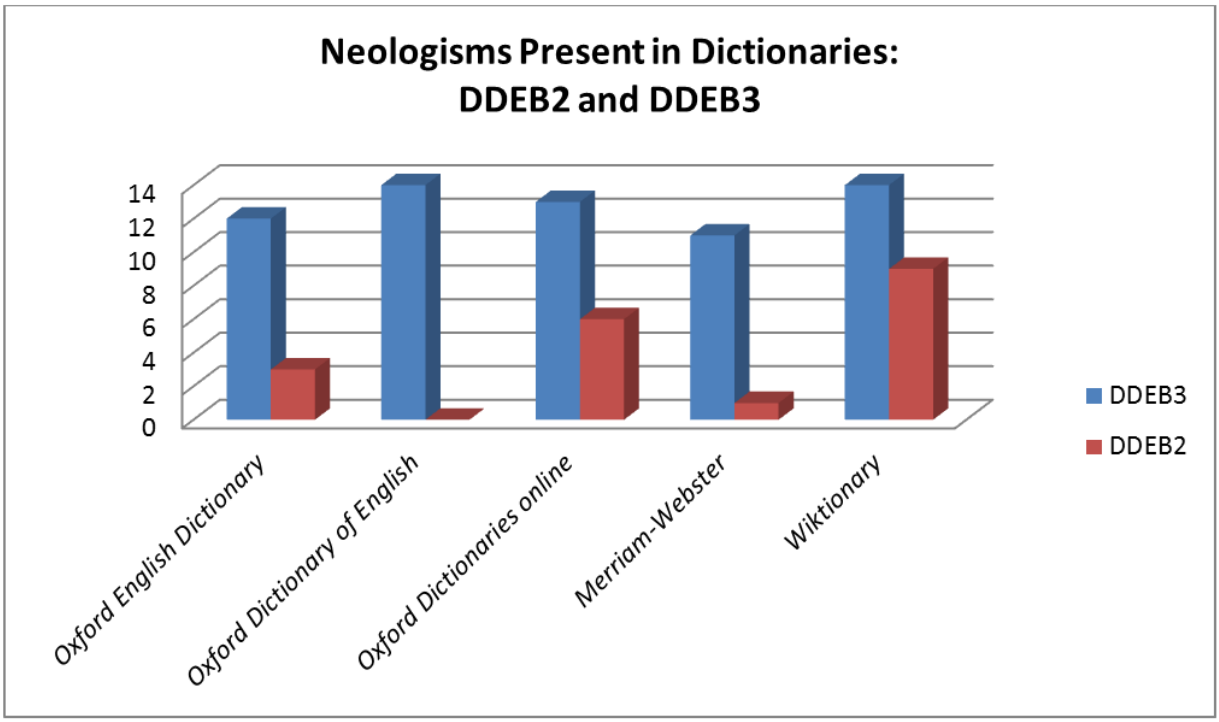


Figure 5.3: Number of neologisms present in dictionaries

As is clear from Figure 5.3, none of the DDEB2 neologisms appear in the *Oxford Dictionary of English* (*ODE)*, the latest printed edition of which was published in 2010. However, six appear in the electronic version of the same dictionary (*ODO*) (see Table 5.1 above). *Merriam-Webster* (*MW*) consistently contains fewer neologisms than the other dictionaries in this study, and indeed performs below the other dictionaries in all

other areas, as will be shown in the following sections. In total, seven of the 15 words in DDEB3 had definitely entered dictionaries before the start of this study (dates having been provided by the dictionaries themselves, although the dates provided by the *Oxford English Dictionary* (*OED*) have, during the course of this project, proved unreliable, so must be viewed with care). These seven were: 'acedia', 'conurbation', 'hubristic', 'promissory note', 'tenebrous', 'upskill' and 'waterboarding'. Four of these words will likely have a long history of use in the media, being the four I termed 'reincarnated' words, since they are believed to have been in regular use, to have fallen out of favour then to have 'risen again' in recent times, resulting in their entry into *Wiktionary* in 2005 and 2006 (see 5.4.1.2). These are 'acedia', 'conurbation', 'hubristic' and 'tenebrous', which entered *OED* many years ago. 'Warrantless' and 'promissory note' may also have a longer history of use than some of the other neologisms in DDEB3; the former entered *OED* in 1921, and the latter has no known date. 'Upskill' entered *OED* in 1993 and therefore is likely to fall between these latter neologisms and the remaining eight.

5.3.2 Dictionary Entry Components

In this section I present the results of comparing the 24 components (standardised and non-standardised) (see 3.4.3) found in entries in the five dictionaries. I examine the numbers and types of these components, as part of my wider investigation of the degrees of comprehensiveness enjoyed by each dictionary. In particular, I compare collaborative dictionary *Wiktionary*'s approach to components with that of the expert-produced dictionaries (looking at both numbers of components and more importantly quality) and I discuss how this affects the overall level of detail in its entries. In the course of this, I lay the groundwork for comparisons of neologism definitions (including differing defining styles (Atkins and Rundell 2008: 450-52) in different dictionaries (5.3.4), and I explain what has been learnt about *Wiktionary*, in response to Research Question 1, shown above.

In considering these issues, I bear in mind the differences in neologisms from the two datasets, Dictionary Date of Entry Batch 1+2 (DDEB1+2) and DDEB3, which indicate how long a word has been in dictionaries, particularly *Wiktionary*, and thus its position

within the neologic life-cycle. I also consider dictionary entries based on their relationship to corpora: 'corpus-based', 'corpus-informed' or 'collaborative'.

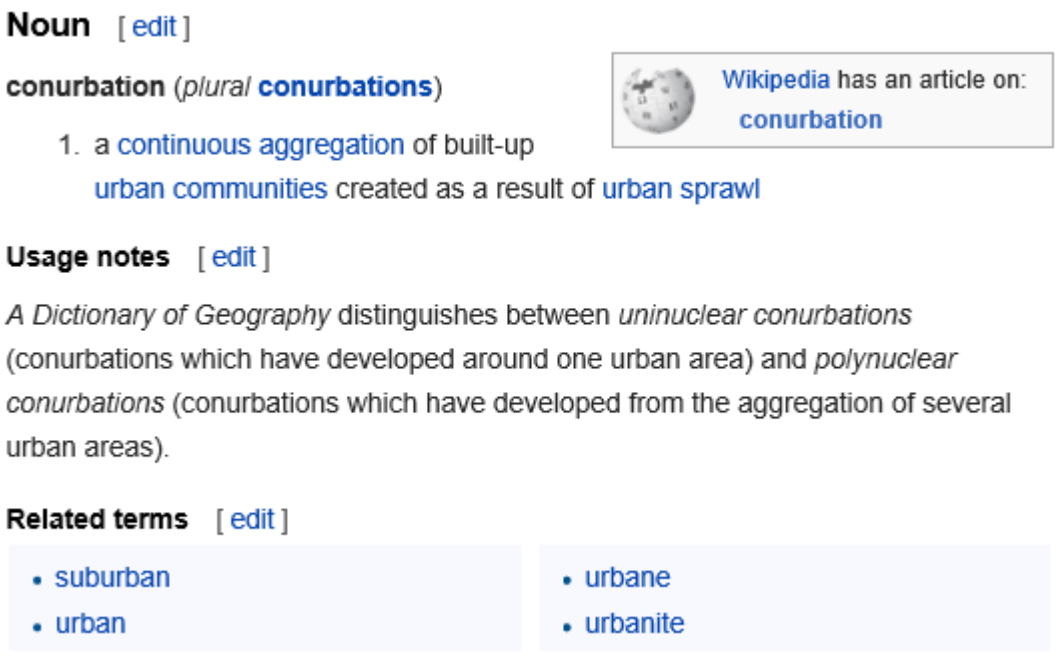
Most of the entries discussed here are from DDEB3 (having entered dictionaries between 2000 and 2008) since those in DDEB2 (entering between 2008 and 2014) were still very simple. This suggests that entries do gain additional information and components over time, a process which was shown by this study to be more prevalent, extensive and transparent in *Wiktionary*. One particularly useful (non-standardised) component present in *Wiktionary* but none of the expert-produced dictionaries was the full Contents Panel (with hyperlinks) at the beginning of most entries, to help the user to move around the entry more quickly. This is a useful feature for long entries such as that for 'acedia'[107] (DDEB3), and is an example of the way *Wiktionary* offers more detailed entries than the expert-produced ('corpus-based' / 'corpus informed') dictionaries in the study. *Wiktionary* also carries translations into languages as diverse as Afrikaans, Japanese and French, adding further detail not available in the non-collaborative dictionaries.

Changes over time in the expert-produced dictionaries tend to happen across the website (for example the addition of frequency information to all entries) while in *Wiktionary* they happen to individual entries, depending on which contributors have been involved. This can result in a lack of balance in the development of *Wiktionary* new word entries, with some growing in complexity (in terms of the number and quality of components) more quickly than others, and to a greater extent. This must be considered one of the 'down-sides' of collaborative dictionary-making; the lack of editorial oversight means that, for example, some new entries can be overlooked if no-one is 'championing' them. Thus in nearly six years no additional information was added to the original entry for 'wellderly', whereas as is shown in 5.3.4, many additional components were added to the entry for 'promissory note' over the same time period.

In order to further address Research Question 1, I specifically examine the issue of detail in dictionary entries. The 24 dictionary components were compared quantitatively as well as qualitatively. The qualitative assessment involved subjectively

[107] https://en.wiktionary.org/wiki/acedia

examining the quality of the components in terms of the amount and quality of information they convey. This was by far the more important of the two comparisons in this largely qualitative lexicographical exploration, since a poor quality component can be more damaging, in terms of confusion to the reader and negative impact on the reputation of the publication, than the absence of the component altogether. One such example of this is the 'usage note' attached to *Wiktionary*'s entry for 'conurbation' (see Figure 5.4). Usage notes generally provide advice to the reader on how a word should be used or what pitfalls to avoid (Atkins and Rundell 2008: 233). *Wiktionary's* entry guidelines agree, stating that Usage Notes should 'describe how a word is used' (Wiktionary 2016c). Here, however, the usage note is simply providing variations on the term 'conurbation', with no explanation of how they are used. It appears that the contributor who added this Usage Note to the entry in July 2008[108] (who was later sanctioned for inappropriate behaviour[109]) failed to follow the guidelines, and no-one has since corrected the error.



Noun [edit]

conurbation (*plural* **conurbations**)

1. a continuous aggregation of built-up urban communities created as a result of urban sprawl

Wikipedia has an article on: **conurbation**

Usage notes [edit]

A Dictionary of Geography distinguishes between *uninuclear conurbations* (conurbations which have developed around one urban area) and *polynuclear conurbations* (conurbations which have developed from the aggregation of several urban areas).

Related terms [edit]

- suburban
- urban
- urbane
- urbanite

Figure 5.4: Usage note for the noun 'conurbation' in *Wiktionary*[110]

[108] https://en.wiktionary.org/w/index.php?title=conurbation&oldid=4802590
[109] See https://en.wiktionary.org/w/index.php?title=User:G1257&action=edit&redlink=1
[110] https://en.*Wiktionary*.org/wiki/conurbation

It seems that the terms 'uninuclear conurbation' and 'polynuclear conurbation' would be more appropriate as 'related terms', although this would require the creation of separate pages for each of them, since all related terms must be 'wikified', that is, have a wiki entry of their own (whether that be in *Wiktionary*, *Wikipedia* or any other associated page) (Wiktionary 2016c).

The Usage Note for 'mitigate' in *Oxford Dictionaries* online (*ODO)* in Figure 5.5 is much more appropriate.

Usage

The verbs mitigate and militate do not have the same meaning, although the similarity of the forms leads many people to confuse them. Mitigate means 'make (something bad) less severe', as in drainage schemes have helped to mitigate this problem, while militate is nearly always used in constructions with against to mean 'be a powerful factor in preventing', as in these disagreements will militate against the two communities coming together

Figure 5.5: Usage note for the verb to 'mitigate' in *Oxford Dictionaries* online[111]

Here the reader is warned against confusing 'mitigate' with 'militate', is given the meanings of the two terms and is provided with examples of how each should be used correctly. This is what we would expect from a standard Usage Note.

Despite the obvious importance of qualitative comparisons of dictionary components, it was still necessary to consider the number of elements present/absent in dictionary entries, especially where a component appeared in only one dictionary, and added significant value to that entry. The quantitative assessment involved looking at which neologism attracted the most components (standardised and non-standardised), which dictionary tended to include the most components, and whether the number of components incorporated into a dictionary entry changed over time. The last of these was only fully possible in *Wiktionary*, due to its 'Revision History' for each entry showing every change ever made to the page. However to a lesser degree it was possible to get an idea of whether new components tended to be added to entries in the other dictionaries, using screenshots taken of the entries at the cut-off point for data collection in August 2014, and their corresponding entries in October 2016. Changes to

[111] https://en.oxforddictionaries.com/definition/mitigate

the *Oxford Dictionary of English* between its three main editions (1998, 2003 and 2010) were easy to identify in hard copy. It was difficult to identify changes to *ODO,* since the website has been relaunched since 2014. However entries in *OED* now include a number of additional components that were not present two years ago, as is mentioned below.

When we compare the total number of standardised and non-standardised components (across all neologisms) used in the combined entries for each dictionary in DDEB2 and DDEB3, in both cases, in answer to Research Question 1, we learn that, as we might expect from *Wiktionary*'s less formal approach to style its entries contain significantly more components than any other dictionary (see Figure 5.6).

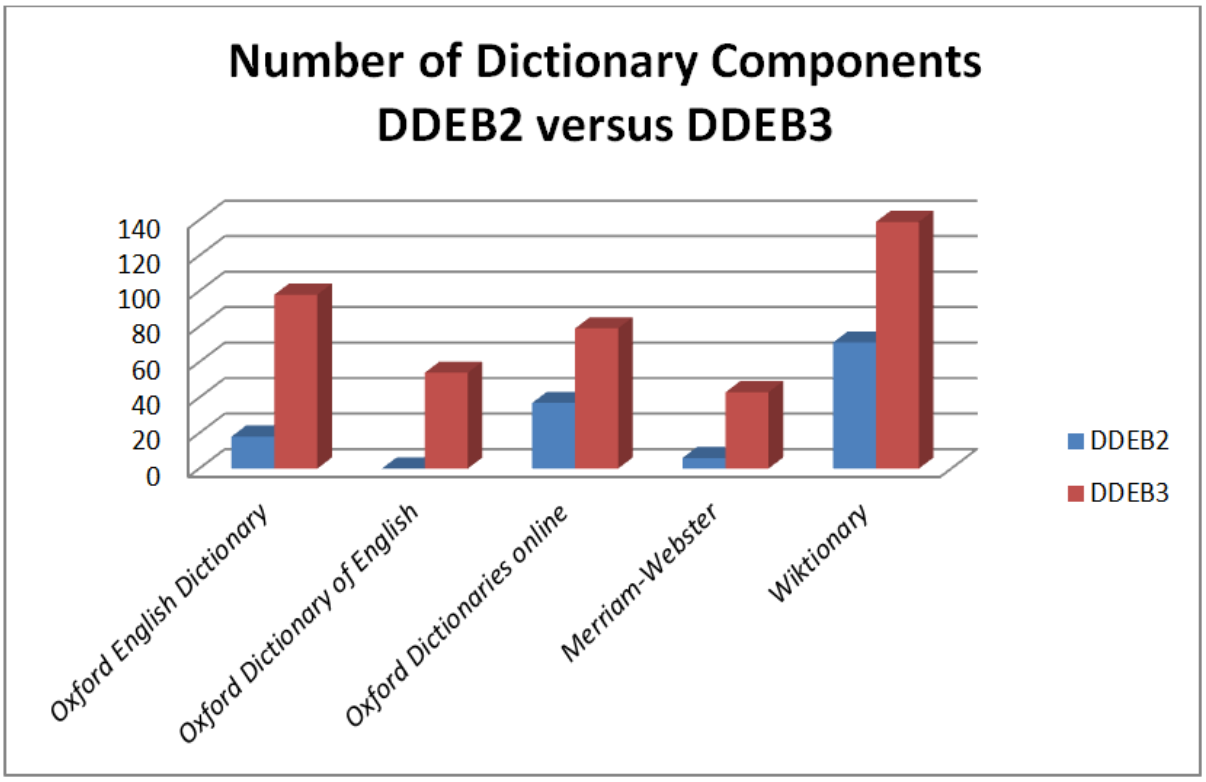


Figure 5.6: Comparison of dictionary components in all neologisms across datasets DDEB2 and 3

For DDEB2 *Wiktionary* is followed by *Oxford Dictionaries* online (*ODO*), and for DDEB3 *Wiktionary* is followed by the *Oxford English Dictionary* (*OED*). While this finding should be tempered by consideration of the quality of those components, it is still interesting

to note that *Wiktionary* components are created by (unskilled) collaborative contributors (with no lexicographical resources) working together (Meyer and Gurevych 2012: 271), as opposed to a single or team of lexicographers who have access to more information to guide their development of the entry (for example the *Oxford English Corpus* (OEC)).

Merriam-Webster consistently carries the fewest components: six and 43 across all neologisms in DDEB2 and 3 respectively, as compared with *Wiktionary*'s 71 and 139. It contains no additional information aside from the basics: definitions, sense distinctions, word class markers, grammatical labels, pronunciation guidance and sound files. This is surprising and as yet unexplained.

The same result is found when comparing *Wiktionary* with the expert dictionaries organised into 'corpus-based' and 'corpus-informed', as Figure 5.7 shows.

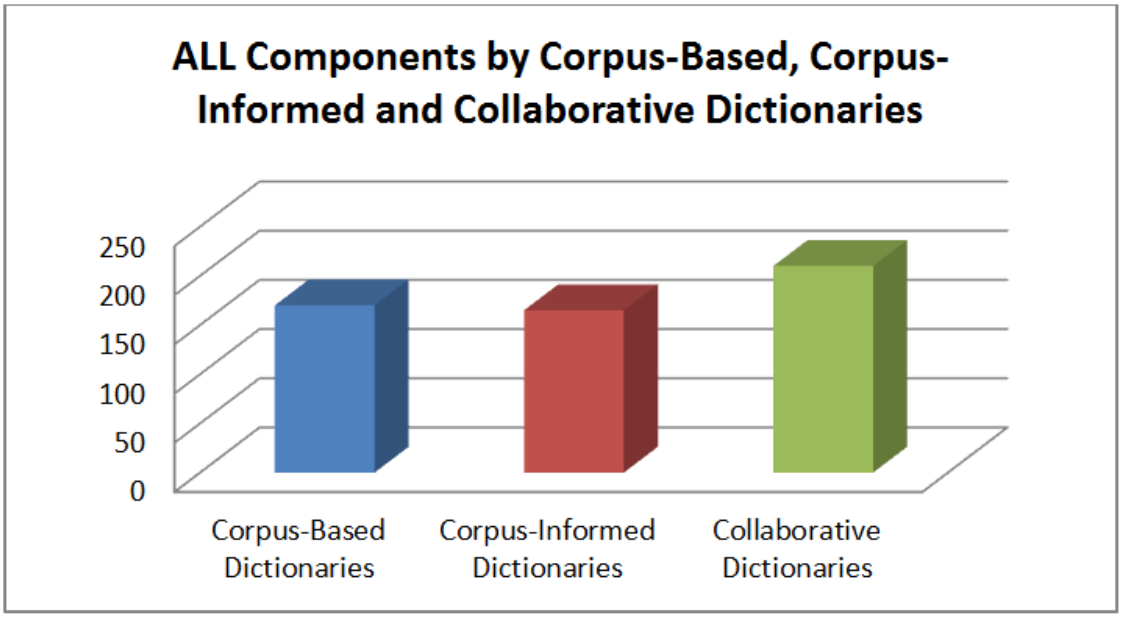


Figure 5.7: Number of components present in dictionaries organised by relationship to corpora

Examining the number of components appearing in each of the dictionary types then, we can conclude that *Wiktionary*, representing collaborative dictionaries, contains more detail than either of the other groups. The higher numbers of dictionary components

achieved by *Wiktionary* is in part due to the fact that it is not limited to including only standardised components, but instead is free to include additional non-standard ones as well (generally not found in other dictionaries), as shown in Table 5.5.

Dictionary Component	**Description**
Standard Dictionary Components	
Headword (lemma)	Indicator of how a word is written
Lexical unit	Subdivisions of headwords (senses)
Menu	List of lexical units
Definition	Explanation of the meaning of a headword
Pronunciation	Guidance on how a word should be pronounced
Etymology	Origin of a word
Spelling variant	Variations in spelling of a word
Word class	Part of speech
Grammar label	Indicator of grammatical information on the headword
Register/style/attitude labels	Indicators of the type of word
Domain label	Marker of the field to which the headword applies
Region label	Indicator of where the word is generally used
Example	Text elucidating the meaning, illustrating use or attesting to the presence of a headword in the language
Usage note	Notes giving additional information on using a headword
Cross-reference	Indicator that more information is available elsewhere
Run-on	Indicators of words derived from the headword
Non-Standard Dictionary Components (mainly used in *Wiktionary*)	
Inclusion date	Indicator of when the word first entered the dictionary (also provided for some words in *OED*)
Revision History	Save-by-save record of every change ever made to an entry
Discussion Forum	Online spaces for discussion of dictionary entries, specifically Talk pages and the Tea Room
Audio File	Sound file added to help with pronunciation (now found in many electronic expert-produced dictionaries)
Translation	Headwords provided in multiple languages (*Wiktionary* only)
Derivative	Marker that the neologism under study derives from another headword, for example in *OED* 'cyberbullying' is a derivative of 'cyber'
Related term	Indicator that a word is linked to the headword, although it is not an actual run-on. In standardised dictionaries this might appear as a Usage Note, however it is treated separately here as *Wiktionary* treats it as a separate element
Synonym	Word which means the same as the headword. Also often included in Usage Notes, but separated here for the same reason as related terms
Contents navigation panel	Panel of hyperlinks to help users move around longer entries in *Wiktionary*

Table 5.5: Standard and non-standard dictionary components (in the context of this study)

The number of components, as shown in Table 5.5, is not the only issue however; as well as the quality of those components we must bear in mind that because *Wiktionary* entries are created collaboratively, any number of people can add elements to them, building the entry up over time. It is likely that this is why *Wiktionary* shows more components than its competitors in Figures 5.7 and 5.8

The information on levels of detail in *Wiktionary* new word entries (gathered during this study in order to answer Research Question 1) is clearly demonstrated when we explore the entries for 'frenemy' (DDEB3), which in total contain the greatest number of dictionary components, closely followed by 'tenebrous' and 'conurbation', as Figure 5.8 demonstrates:

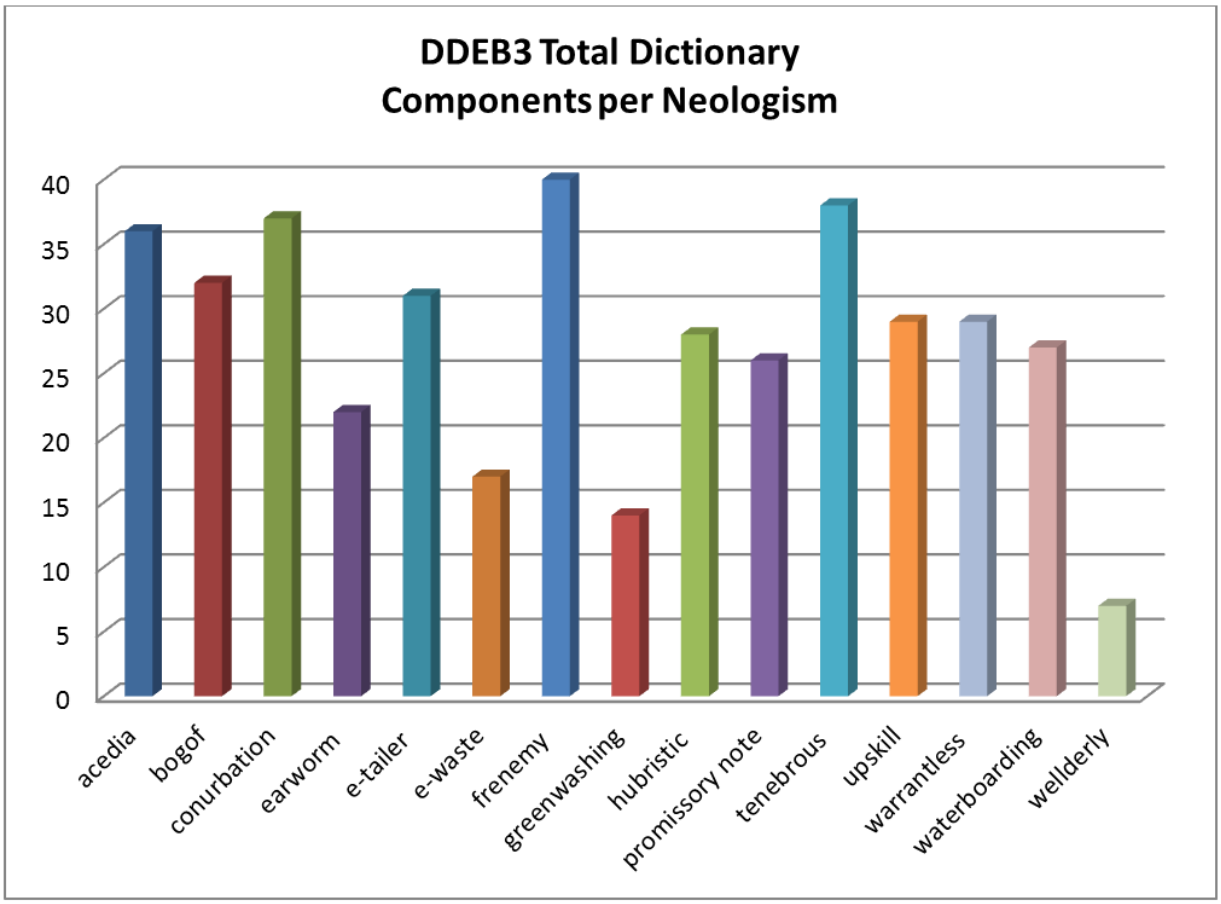


Figure 5.8: Number of components displayed in entries for DDEB3 neologisms across all dictionaries

It is surprising to find ‘frenemy’ in the top spot, since its performance throughout the rest of this study has been unremarkable. It is, however, one of only three neologisms (‘frenemy’, tenebrous’ and ‘warrantless’) in DDEB3 to have entries which include examples in three of the five dictionaries (*Wiktionary*, *ODE* and *ODO*). The examples in *Wiktionary* and *ODE* are attestational and illustrative (carrying date and source information to prove use of the word in the lexicon at large), whereas in *ODO* they are purely illustrative (Atkins and Rundell 2008: 453-4).

Forming part of DDEB3, ‘frenemy’ does not achieve particularly high numbers of neologism appearances in any of the newspapers (see 5.4.2), and it falls in the middle of the dictionary inclusion periods, entering in *Wiktionary* in 2005, *OED* in 2008 and *ODE/ODO* in 2010. However the quality of components used in dictionary entries for ‘frenemy’ is largely high, as shown in Figures 5.9 – 5.14.

Entry Discussion Citations Read Edit History Search

frenemy

Archived revision by DPMaid (talk | contribs) as of 21:14, 22 August 2014.
(diff) ← Older revision | Latest revision (diff) | Newer revision → (diff)

Contents [hide]
1 English
1.1 Alternative forms
1.2 Etymology
1.3 Pronunciation
1.4 Noun
1.4.1 Synonyms
1.4.2 Translations
1.5 See also

English

A user suggests that this entry be cleaned up, giving the reason: "The quotations, in particular, need clean-up.".
Please see the discussion on **Requests for cleanup**(+) or the talk page for more information and remove this template after the problem has been dealt with.

Alternative forms

- frienemy

Etymology

Blend of *friend* + *enemy*. Likely to have been invented independently multiple times.

Pronunciation

- IPA(key): /frɛ.nɪ.mi/

Noun

frenemy (*plural* **frenemies**)

1. (*humorous*) Someone who pretends to be your friend, but is really your enemy. [quotations ▼]
2. (*humorous*) A fair-weather friend who is also a rival.

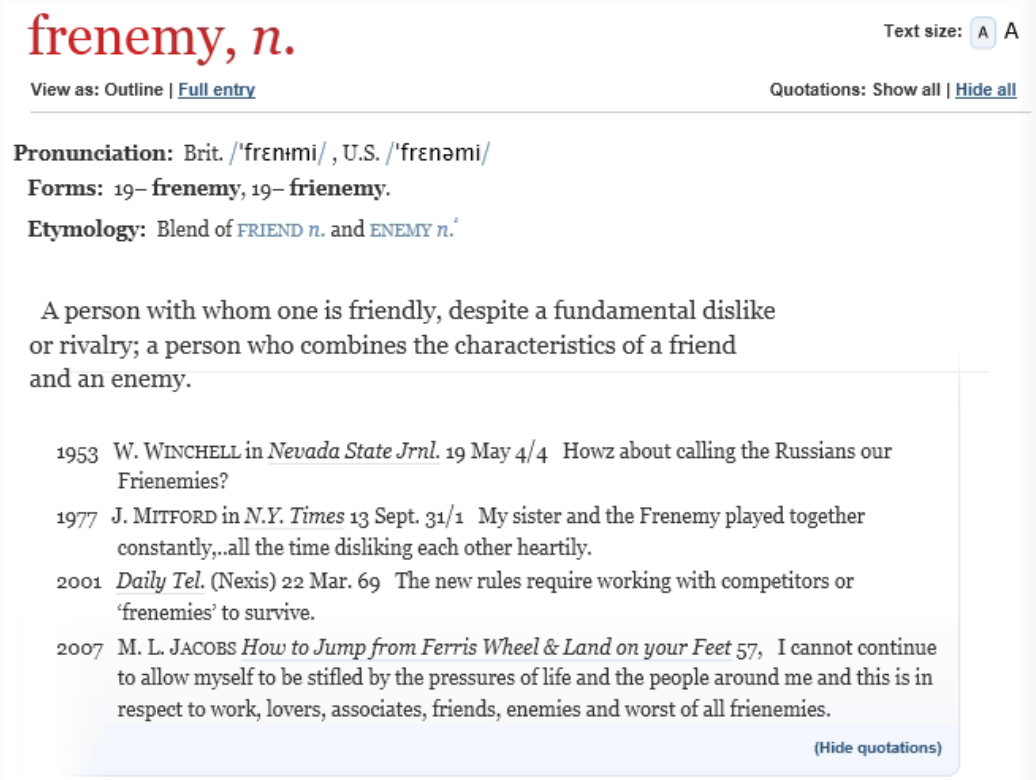
Synonyms

- betrayer
- double-crosser
- traitor
- palhole

Translations

enemy pretending to be a friend [show ▼]

Figure 5.9: *Wiktionary* 2014 entry for 'frenemy'



frenemy, *n.*

Text size: A A

View as: Outline | Full entry

Quotations: Show all | Hide all

Pronunciation: Brit. /ˈfrɛnɪmi/ , U.S. /ˈfrɛnəmi/

Forms: 19– **frenemy**, 19– **frienemy**.

Etymology: Blend of FRIEND *n.* and ENEMY *n.*

A person with whom one is friendly, despite a fundamental dislike or rivalry; a person who combines the characteristics of a friend and an enemy.

1953 W. WINCHELL in *Nevada State Jrnl.* 19 May 4/4 Howz about calling the Russians our Frienemies?

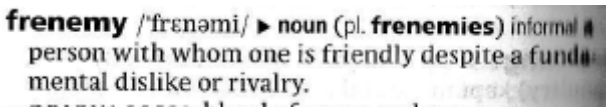
1977 J. MITFORD in *N.Y. Times* 13 Sept. 31/1 My sister and the Frenemy played together constantly,..all the time disliking each other heartily.

2001 *Daily Tel.* (Nexis) 22 Mar. 69 The new rules require working with competitors or 'frenemies' to survive.

2007 M. L. JACOBS *How to Jump from Ferris Wheel & Land on your Feet* 57, I cannot continue to allow myself to be stifled by the pressures of life and the people around me and this is in respect to work, lovers, associates, friends, enemies and worst of all frienemies.

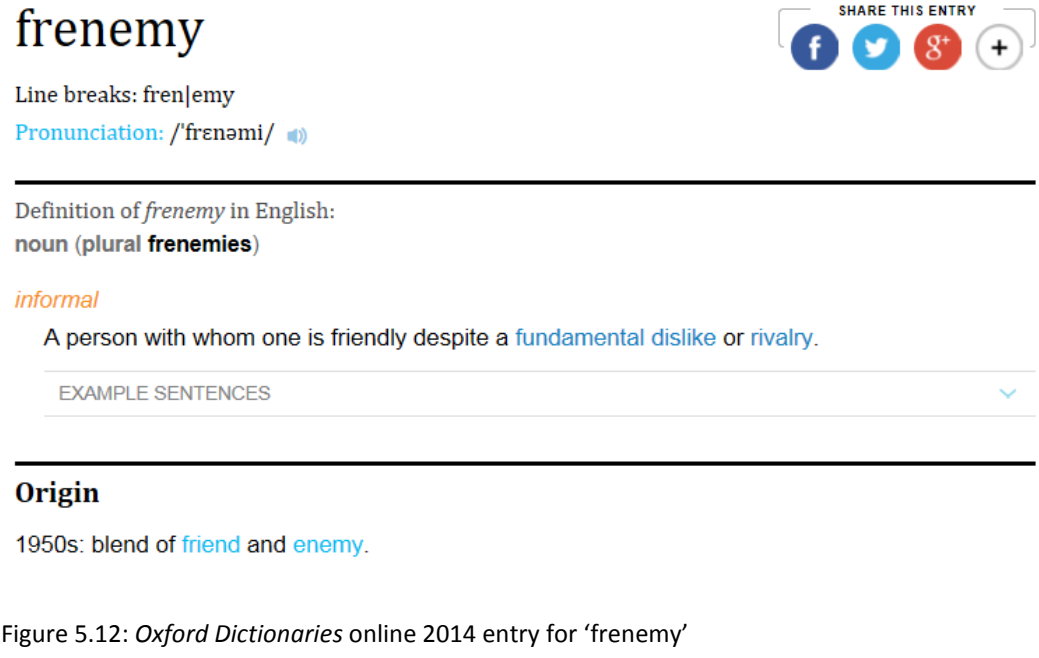
(Hide quotations)

Figure 5.10 *Oxford English Dictionary* 2014 entry for 'frenemy'



frenemy /ˈfrɛnəmi/ ▸ noun (pl. **frenemies**) informal a person with whom one is friendly despite a fundamental dislike or rivalry.

Figure 5.11: *Oxford Dictionary of English* 2014 entry for 'frenemy'



frenemy

SHARE THIS ENTRY

Line breaks: fren|emy

Pronunciation: /ˈfrɛnəmi/

Definition of *frenemy* in English:

noun (**plural frenemies**)

informal

A person with whom one is friendly despite a fundamental dislike or rivalry.

EXAMPLE SENTENCES

Origin

1950s: blend of friend and enemy.

Figure 5.12: *Oxford Dictionaries* online 2014 entry for 'frenemy'

'Advertising mogul Sir Martin Sorrell used the term "frenemy" to describe the phenomenon.'

'Rob says that, at best they should be understood as our `frenemy '.'

'I was very happy it was not someone I respected," she said of her frenemy, whom she called a "coward."'

'Rob called the company a frenemy of the internet generation.'

Figure 5.13: Example sentences for 'frenemy' appearing in *Oxford Dictionaries* online

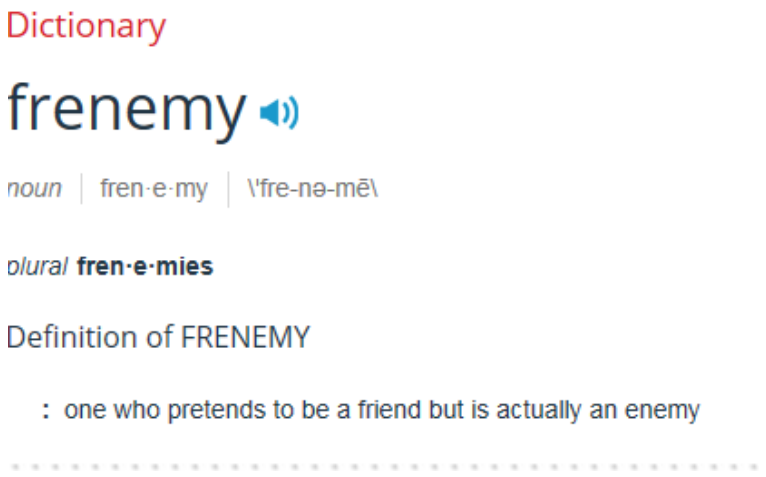


Figure 5.14: *Merriam-Webster* 2014 entry for 'frenemy'

Entries for 'frenemy' utilise 15 of the 24 dictionary components identified in this study, 10 of which are standardised elements as explained by Atkins and Rundell (2008: 385-462) and five of which are non-standard. As shown in Table 5.6, the spread of these components is broad, however only *Wiktionary* and *Oxford Dictionaries* online include components not used by any other dictionary. These additional components add to the level of detail that *Wiktionary* can offer to entries for 'frenemy', compared with the expert-produced dictionaries. Returning to Research Question 1, we can also see how the fact that these entries can be added to by a range of different contributors – each following the less formal inclusion and style guidelines of the site and each updating the site every time they save a changed entry – demonstrates the level of responsiveness offered by *Wiktionary*, an advantage unmatched by any of the expert-produced dictionaries.

Dictionary Component	Dictionary
Definition	OED, ODE, ODO, MW, W
Date of first inclusion*	OED, W
Word class	OED, ODE, ODO, MW, W
Multiple senses	W
Grammar label	ODE, ODO, W
Pronunciation guidance	OED, ODE, ODO, MW, W
Register/style/attitude label	ODE, ODO, W
Sound file*	ODO, MW
Spelling variations	OED, W
Examples/quotations/citations	OED, ODO, W
Synonyms*	W
Morphology	OED, ODO, W
Etymology	ODO
Translation*	W
Contents Panel*	W

*Non-standardised components (based upon Atkins and Rundell 2008)
Table 5.6: Dictionary components and the dictionaries in which they appear for 'frenemy'

All of the definitions for 'frenemy' are of the classical 'genus-differentiae' model, in which a superordinate term positions the headword within the correct semantic category, the former being a person, and the latter some variation on 'pretends to be a friend but is really an enemy'. This is a popular defining strategy and works well in this case (Atkins and Rundell 2008: 414). The use of differing defining strategies for the neologisms under study here is discussed in full in 5.3.4.

All of the entries which present morphology for 'frenemy' agree that it is a blend of 'friend' and 'enemy', and *ODO* claims that it in fact first appeared in the 1950s. The two alternative spellings agree on use of the correct form of the first half of the blend ('frien-' instead of 'fren-') and all of the grammar labels relate to pluralisation of the term, since it follows an irregular pattern ('+ies' rather than '+s'); this is as per the second half of the blend, 'enemy'.

OED, *ODE* and *ODO* agree on the pronunciation of 'frenemy' (although *MW* and *Wiktionary* offer slightly different guidance). Atkins and Rundell state that both the International Phonetic Alphabet (IPA) and Speech Assessment Methods Phonetic Alphabet (SAMPA) are accepted forms of pronunciation guidance (2008: 206), however it is not known what method *MW* uses, and *Wiktionary*'s entry guidance[112] makes no

[112] https://en.Wiktionary.org/wiki/Wiktionary:Pronunciation

mention of SAMPA. From reference to a SAMPA chart[113], I would argue that *Wiktionary*'s pronunciation guidance is based on this system, but MW seems to use an entirely bespoke system. In general, provision of pronunciation guidance is sporadic at best across the five dictionaries, as Tables 5.7 and 5.8 show.

	Dictionary				
Neologism	*OED*	*ODE*	*ODO*	*MW*	*W*
bankster	N/A	N/A	✓	N/A	✓
cyberbullying	x	N/A	✓	✓	N/A
cyberchondriac	x	N/A	✓	N/A	x
diabesity	N/A	N/A	✓	N/A	x
floordrobe	N/A	N/A	N/A	N/A	✓
gendercide	N/A	N/A	N/A	N/A	x
globesity	N/A	N/A	N/A	N/A	x
hyperlocal	N/A	N/A	✓	N/A	x
rewilding	x	N/A	N/A	N/A	N/A
sovereign debt	N/A	N/A	N/A	N/A	x
superphone	N/A	N/A	N/A	N/A	x

Table 5.7: Provision of pronunciation guidance in DDEB2 neologisms, by dictionary

	Dictionary				
Neologism	*OED*	*ODE*	*ODO*	*MW*	*W*
acedia	✓✓	✓	✓	✓	✓
bogof	✓✓✓	✓	✓	N/A	✓
conurbation	✓	✓	✓	✓	✓✓
earworm	N/A	x	✓	✓	✓
e-tailer	✓✓	x	✓	x	x
e-waste	N/A	x	x	✓	x
frenemy	✓✓	✓	✓	✓	✓
greenwashing	✓✓	x	N/A	✓	N/A
hubristic	✓	✓	✓	✓	x
promissory note	✓	x	✓	x	x
tenebrous	✓	✓	✓	✓	✓✓✓
upskill	x	x	✓	N/A	x
warrantless	x	x	✓	N/A	✓
waterboarding	✓✓	x	✓	✓	x
wellderly	N/A	N/A	N/A	N/A	x

Table 5.8: Provision of pronunciation guidance in DDEB3 neologisms, by dictionary

[113] https://en.wikipedia.org/wiki/Speech_Assessment_Methods_Phonetic_Alphabet_chart_for_English

As these tables show, *Wiktionary* is poor in providing pronunciation guidance on the newer neologisms, however it is much more successful on the older neologisms, demonstrating that these entries have been expanded over time, through the actions of contributors (see 5.3.3 for a full discussion of this issue). However in DDEB3, *ODE* performs particularly badly. (On both tables, 'N/A' indicates that the word is not included in that dictionary.) In *OED* and, to a lesser extent, *Wiktionary*, multiple pronunciation forms are often shown; these are generally British English versus American English, although sometimes there is an additional option for one or other of these. There is sometimes disagreement over the correct pronunciation of a word between these two English varieties, however. For example in my view the English and US versions for 'frenemy' in *OED* are the wrong way around. *ODE* and *ODO* appear to agree with me, using what *OED* calls the US pronunciation in their (British) IPA. *Wiktionary* agrees with *OED*, but utilises a slightly odd version of what appears to be IPA, with unusual syllable/stress markers, see Figures 5.15 – 5.19. This can be confusing for users who are accustomed to seeing IPA in, for example, Oxford dictionaries.

Figure 5.15: *OED* pronunciation guidance for 'frenemy', featuring British and American English pronunciation

Figure 5.16: *Wiktionary* pronunciation guidance for 'frenemy'

As we can see from the two figures above, the ostensibly British pronunciation in *Wiktionary* actually matches the American pronunciation in *OED.*

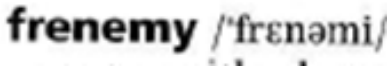


Figure 5.17: *ODE* pronunciation guidance for 'frenemy'

Figure 5.18: *ODO* pronunciation guidance for 'frenemy'

Figure 5.19: *MW* pronunciation guidance for 'frenemy'

We can see from Figures 5.17 and 5.18 that *ODE* and *ODO* use IPA transcriptions which *OED* attributes to US English, while in Figure 5.19, *MW* uses a slightly different system, identifiable by the final symbol which appears to have something akin to an umlaut above the 'e'. Surprisingly only *ODO* and *MW* offered sound files for 'frenemy' at this time. Indeed in both cases, every entry with pronunciation guidance also included a sound file. Both recordings were accurate (both sites have since been updated, and the recordings may have been replaced). At the time of data collection (August 2014 onwards) *OED* did not include sound files as standard across all its entries, however it also has since been updated, and now does so. Where *Wiktionary* included a sound file, it also included a panel giving information on who uploaded the sound file and when. However only two out of 23 neologism entries include sound files.

The synonyms provided by *Wiktionary* for 'frenemy' are not exact equivalent terms, but they are close enough to be considered acceptable. As Atkins and Rundell point out, 'true synonyms are extremely rare, if they exist at all' (2008: 135). None of the other dictionaries provide synonyms for these words, probably because they are too new to have established words to which they can be considered sufficiently close in meaning. *ODE*, *ODO* and *Wiktionary* all carry register/style/attitude labels for 'frenemy' which are accurate, although different; two are 'informal', and *Wiktionary*'s is 'humorous', which is a less common label among these neologisms in these dictionaries. (As noted in 3.4.3, labelling is a slightly less standardised component than the others discussed here; as a consequence and in order to avoid inconsistences, for the purposes of this study I group all three label types together.) Indeed within DDEB2 and 3, register/style/attitude labels are surprisingly scarce in the expert-produced dictionaries; *Merriam-Webster*, for example uses none at all, and the highest usage, in *ODO*, still totals only six across the

26 neologisms. The presence of more labels in *ODO* than other expert-produced dictionaries could be attributed to its 'corpus-based' status. However it would be possible for 'corpus-informed' dictionaries such as the *OED*[114] to draw information from the *Oxford English Corpus* (*OEC*[115]) for labels such as register, yet 'corpus-informed' dictionaries were found to perform very poorly in this regard, with the *OED* responsible for just two labels, and MW using no labelling at all.

'Corpus-based' dictionaries then are responsible for nine register/style/attitude labels, and collaborative dictionaries are responsible for six (although there is only one dictionary in this category, while there are two in each of the others, meaning that *Wiktionary* has performed slightly better than the 'corpus-based' dictionaries). 'Corpus-informed' dictionaries are responsible for just three. These labels are spread across entries for nine of the 26 neologisms in the five dictionaries, with 'frenemy' containing labels in three of them. Table 5.9 shows the distribution of register/style/attitude labels across the dictionaries.

Neologism	Dictionary Date of Entry Batch	Dictionary(ies)	Register/Style/Attitude Label
acedia	DDEB3	*ODE*	Literary
		ODO	Literary
bankster	DDEB2	W	Informal, Derogatory
		ODO	Derogatory
cyberchondriac	DDEB2	*OED*	Depreciative
diabesity	DDEB2	*ODO*	Informal
floordrobe	DDEB2	W	Humorous
frenemy	DDEB3	W	Humorous
		ODE	Informal
		ODO	Informal
superphone	DDEB2	W	Informal
tenebrous	DDEB3	*ODE*	Literary
		ODO	Literary
warrantless	DDEB3	*OED*	Rare

Table 5.9: Register markers across neologisms by dictionary

Four of the neologisms in Table 5.9 are marked 'informal': 'bankster', 'diabesity', 'frenemy' and 'superphone'. It may be that these register markers owe more to the

[114] See for example https://www.oxforddictionaries.com/news-and-press/oxford-dictionaries-faq
[115] https://en.oxforddictionaries.com/explore/oxford-english-corpus

neologisms' 'new' status than to the words themselves, suggesting that all new words begin life in 'informal' usage. Other labels allow lexicographers to provide additional information about a word that is not contained elsewhere in the dictionary entry, for example that a word is archaic or humorous (Atkins and Rundell 2008: 229). We would not expect to see 'derogatory' as a label if the definition read 'a derogatory term for …'. (Even in *Wiktionary*, we would expect fellow contributors to notice the duplication and take action to fix it (Meyer and Gurevych 2012: 271).) Indeed it is perhaps surprising that not only *Wiktionary*, but also *ODO* apply a 'derogatory' label to entries for 'bankster', since in each case this impression is supplied by the definition, as shown in Figures 5.20 and 5.21.

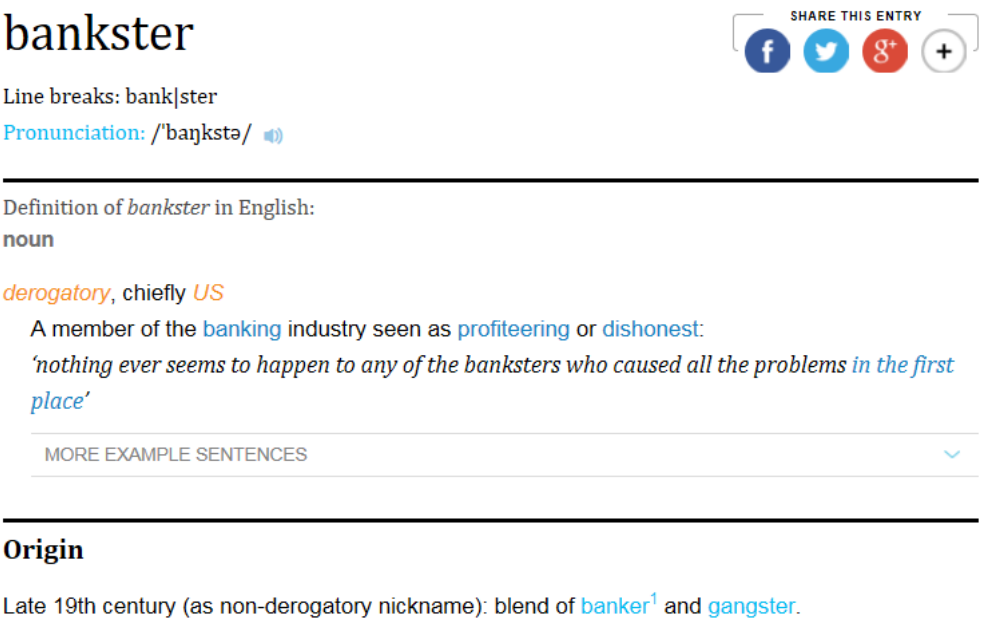


Figure 5.20: *ODO* entry for 'bankster'

bankster

Archived revision by IjonTichyIjonTichy (talk | contribs) as of 00:05, 19 August 2014.
(diff) ← Older revision | Latest revision (diff) | Newer revision → (diff)

Contents [hide]
1 English
1.1 Etymology
1.2 Pronunciation
1.3 Noun
1.3.1 Translations
1.4 References

English

Etymology

Blend of *banker* + *gangster*

Judge Ferdinand Pecora coined the term Bankster. In June 1933, his image appeared on the cover of Time magazine, seated at a US Senate table, a cigar in his mouth. Pecora's hearings had coined a new phrase, "banksters" for the finance "gangsters."

The term was later used by Léon Degrelle, Belgian fascist politician and journalist, in 1937 as a pejorative term for high financiers.

Pronunciation

- (*UK*) IPA(key): /ˈbæŋkstə/

Noun

bankster (*plural* **banksters**)

1. (*informal*, *derogatory*) A banker who is seen as criminally irresponsible, or as extorting bailout money from the taxpayers.
[quotations ▼]

Translations

Translations [show ▼]

References

- The man who busted the banksters, *Smithsonian Magazine*

Figure 5.21: *Wiktionary* entry for 'bankster

In both cases, the definition of the word indicates its derogatory nature; the label is largely superfluous. Other labelling of neologisms in these dictionaries appears more appropriate however. 'Cyberchondriac' is labelled 'depreciative' by *OED,* 'acedia' and 'tenebrous' are labelled as 'literary' by *ODE* and 'humorous' labels are applied to 'floordrobe' and 'frenemy' by *Wiktionary*. In each case, the label complements and builds on the information provided in the definition.

In addition to register/style/attitude labels, some of the dictionary entries also carry 'domain' labels indicating the field or context to which it applies (Atkins and Rundell 2008: 227). For example 'diabesity' carries the domain label 'pathology' in *Wiktionary*. 'Region' labels are also supplied, showing where the word is generally used (Ibid). Both of these labels are rare in these neologism entries; there are only four domain labels across all 26 neologisms, and the same for regional labels. Of the latter, one ('bogof') indicates British English usage (in *OED*) and the rest ('warrantless', 'waterboarding' and 'bankster') indicate usage in the United States (the first two in *OED* and the latter in *ODO*). Grammatical labels are more widespread. Most provide the correct plural form, or indicate the nature of a noun, for example countable/uncountable ('mass noun'). In *Wiktionary*, all entries except 'diabesity' include grammatical labels. In the expert-produced dictionaries, a handful of DDEB3 neologisms carry grammatical labels (nine in 'corpus-based' dictionaries and two in 'corpus-informed') while in DDEB2 just two words ('cyberbullying' and 'diabesity') carry grammatical labels, both in *ODO.* In terms of providing additional information to help readers use the neologisms correctly, *Wiktionary* out-performs the expert-produced 'corpus-based' and 'corpus-informed' dictionaries.

It is interesting that several of the DDEB1, 2 and 3 words seem to occur sufficiently frequently in *Oxford English Corpus* (*OEC*) to merit an entry in the 'corpus-based' dictionaries (*ODE* and *ODO*) (see Tables 5.10 and 5.11), including 'predatory lending', which has in fact yet to enter any dictionary. It occurs 406 times in the *OEC*, between 2000 and 2012 (based on research findings derived from the *Oxford English Corpus*, Oxford University Press). This indicates that the word has been in use for a number of years (indeed my own study shows it dating back to 1993 in the *Independent*). It might therefore be suggested that the reason these words have failed to move on to the next stage (presence in an Oxford dictionary), is that they have failed to meet other, perhaps less publicised, inclusion criteria. However a more mundane explanation also exists, that of human error, the term not having been included simply by mistake.

Neologism	Number of Entries in *OEC*
bankster	19
buzz marketing	59
cold peace	63
cyberbullying*	383
cyberchondriac*	8
diabesity	14
floordrobe	4
gendercide	21
globesity	3
hyperlocal*	133
newer markets	36
open education	48
predatory lending	406
rewilding*	60
round pound	10
sodcasting	2
sovereign debt	925
superphone	26
tablet computing	19

Table 5.10: *Oxford English Corpus* information for DDEB1+2 entries (DDEB1 neologisms (in red) not included in any dictionary as at 31 August 2014). (Based on research findings derived from the *Oxford English Corpus*, Oxford University Press)
* Includes spelling variants

Neologism	Number of Entries in *OEC*
acedia	40
bogof	54
conurbation	629
earworm*	266
e-tailer	918
e-waste	289
frenemy	29
greenwashing*	187
hubristic	373
promissory note	852
tenebrous	91
upskill	89
warrantless	876
waterboarding*	2,217
wellderly	2

Table 5.11 *Oxford English Corpus* information for DDEB3 entries. (Based on research findings derived from the *Oxford English Corpus*, Oxford University Press)
* Includes spelling variants

As Tables 5.10 and 5.11 show there appears to be sufficient information in the *OEC* to enable the Oxford dictionaries to provide more detailed entries.

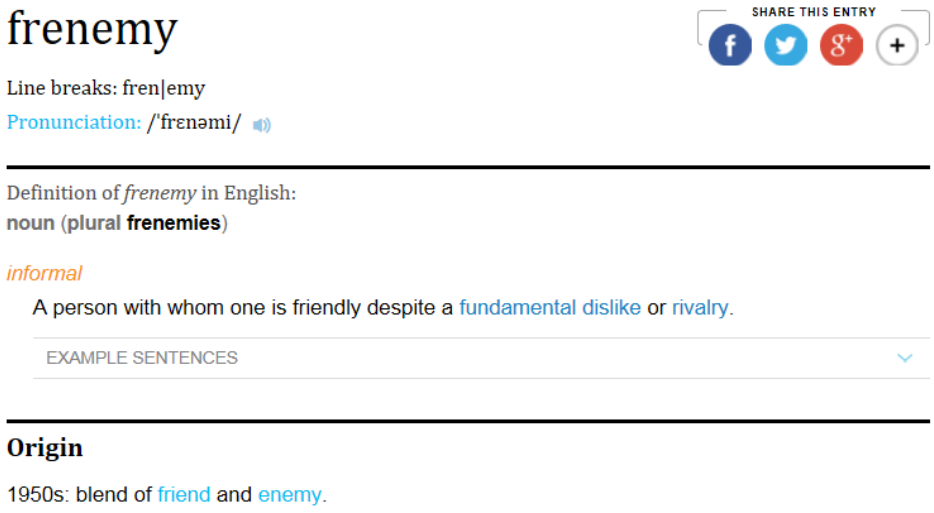


Figure 5.22 *Oxford Dictionaries* online 2014 entry for 'frenemy', featuring the register 'informal' and 'example sentences' accessible via a link

'Frenemy' is one of only five neologisms in *Wiktionary* to carry examples (the term 'examples' here being used to incorporate both standard examples and quotations/citations, since in the debate about standardised dictionary components they are grouped together, then distinguished based upon whether they are 'illustrative (providing information on how the word behaves amongst its peers), 'elucidating' (further explaining the meaning by demonstrating usage) or 'attestational' (proving that the word was in existence) (Atkins and Rundell 2008: 453-4). It also carries examples in *ODO* and *OED*. The spread of examples across the dictionaries, by Dictionary Date of Entry Batch (DDEB) is shown in Tables 5.12 and 5.13.

Dictionary	**Number of Examples/Quotations/Citations out of Total Neologism Entries**
OED	12/12
ODE	0/14
ODO	12/13
MW	0/10
W	5/14

Table 5.12: DDEB3: Number of entries in each dictionary carrying an example/quotation/citation

Dictionary	Number of Examples/Quotations/Citations out of Total Neologism Entries
OED	3/3
ODE	0/0
ODO	6/6
MW	0/1
W	6/9

Table 5.13: DDEB2: Number of entries in each dictionary carrying an example/quotation/citation

Across both of the datasets shown above, all of the *Wiktionary* examples are both attestational and illustrative, as are the majority of the *OED* ones ('waterboarding' and 'conurbation' also contain elucidating examples). All of the *ODO* examples are illustrative. From this we can see that in this context *Wiktionary*'s examples follow the same standardised patterns as expert-produced and 'corpus-based' or 'corpus-informed' dictionaries.

As discussed in 3.4.3, dictionary examples generally come from citation banks (mostly attestational examples used by 'corpus-informed' dictionaries), or from corpora, or they are created by the lexicographer (Atkins and Rundell 2008: 455). This leads to the tension existing between 'authentic' examples which have been used in real-world language, and 'typical' ones which show how people regularly speak: the two are not necessarily the same thing (Fox: 143, 138-9).

We can see the problems of authenticity versus typicality at play in the examples used in entries for these neologisms. The example in Figure 5.23 is one of the examples of 'frenemy' drawn from *ODO*, and Figure 5.24 shows its concordance line from the *Oxford English Corpus* (*OEC*).

'Advertising mogul Sir Martin Sorrell used the term "frenemy" to describe the phenomenon.'

Figure 5.23 *ODO* example for 'frenemy'

Figure 5.24 'Frenemy' concordance from the *OEC*. (Based on research findings derived from the *Oxford English Corpus*, Oxford University Press)

This shows that the example is authentic, but it is a very poor example, failing to fit the requirements of either an illustrative example (providing information on things like register, collocations or syntax), or an elucidating example (complementing the dictionary definition by demonstrating correct usage) (Atkins and Rundell 2008: 454). The other three (illustrative) examples in *ODO* might be more effective, as Figure 5.25 demonstrates.

'Rob says that, at best they should be understood as our `frenemy '.'

'I was very happy it was not someone I respected," she said of her frenemy, whom she called a "coward."'

'Rob called the company a frenemy of the internet generation.'

Figure 5.25: *ODO*'s remaining examples for 'frenemy'

The latter two of these also hail from the *OEC,* the bottom one (British English) from a technology publication and the middle one from the *New York Post* (American English). (Based on research findings derived from the *Oxford English Corpus*, Oxford University Press.)

However two of *Wiktionary's* illustrative (attestational) examples are better.

- **2004**, You know when you dump a guy, only to discover years later that he's evolved into the perfect boyfriend—for the high-school **frenemy** who convinced you to dump him in the first place...? —*The Ex-Factor,* Andrea Semple. [back cover] [2]
- **2005**, So why did we break up? Enter Blaize St. John, **frenemy** extraordinaire. She came, she saw, she stole my boyfriend. —*Single Girl's Guide to Murder,* Joanne Meyer. [back cover] [3]

Figure 5.26: *Wiktionary* examples for 'frenemy'

These examples are simultaneously authentic and typical, and provide much more information on the context in which the word is used and the terms that are used with it than do those drawn from the *OEC*.

It is perhaps surprising that the *ODO* manages to include the examples for 'frenemy' shown above, given that the *OEC*, on which it and the *ODE* both rely for information (both being 'corpus-based') includes only 29 instances of 'frenemy' (based on research findings derived from the *Oxford English Corpus*, Oxford University Press). My own

NTON (*Neologism Tracking in Online Newspapers*) database located 47 instances of the term from a similar period.

The differences between the two *OEC*-based examples above raise an important point. It is my opinion that the *OEC* spreads its net too widely to be an optimum source for dictionary examples. The majority of its publications appear to have been from the USA; the only UK national newspapers I have been able to find in the corpus are *The Telegraph* and *The Guardian*. None of the other newspapers used in the current study seem to have been included in the corpus (the *Independent*, the *Mail* or the *Express*). Of the 47 instances of 'frenemy' found in the *NTON* database, 19 were from *The Guardian*, including three which the *OEC* also picked up (based on research findings derived from the *Oxford English Corpus*, Oxford University Press). 'Corpus-informed' dictionaries such as the *OED* and *MW* can use corpora such as the *OEC* for dictionary components, although neither features any of the *OEC*'s entries as examples for 'frenemy' (indeed MW carries no examples at all across any of the neologisms studied here). *OED* perhaps instead uses a 'citation bank' to generate quotations, as described by Atkins and Rundell (2008: 455).

While *OED* and *Wiktionary* both use attestational examples, the former is constrained by the kind of publications it can use for citations and, by extrapolation, its 'citation bank' (Ibid). *Wiktionary*, meanwhile, adopts more relaxed inclusion criteria (see 3.4.2), and hence the source information, from which examples can be drawn, is more varied. Two of the other *Wiktionary* examples for 'frenemy' are from an album cover and a television programme respectively – see Figure 5.27.

- **1987**, **I Ain't No Joke,** by Eric B. and Rakim, on the album "Paid in Full." "Another enemy / Not even a **frenemy**."
- **2000**, **frenemies** —*Sex and the City*, season 3 episode 16, first aired October 1. [title]

Figure 5.27: *Wiktionary* examples for 'frenemy'

These more ‘popular’ attestation sources can come from any *Wiktionary* contributor wishing to add to the entry, and on returning back to Research Question 1 we can see that this is an example of what we have learnt about *Wiktionary*; that it is consistently more responsive to neologism than are expert-produced dictionaries. This is by virtue of the flexibility offered by its collaborative nature, and its less formal approach to inclusion and style. We also see that it provides a level of detail in its new word entries that is unmatched in any of the other dictionaries.

Although ‘frenemy’ had a very extensive entry in *Wiktionary* in 2014 (shown in Figure 5.28), that had not always been the case, as Figure 5.29 demonstrates.

frenemy

Archived revision by DPMaid (talk | contribs) as of 21:14, 22 August 2014.

(diff) ← Older revision | Latest revision (diff) | Newer revision → (diff)

Contents [hide]

English

A user suggests that this entry be cleaned up, giving the reason: "The quotations, in particular, need clean-up.".

Please see the discussion on **Requests for cleanup**(+) or the talk page for more information and remove this template after the problem has been dealt with.

Alternative forms

- frienemy

Etymology

Blend of *friend* + *enemy*. Likely to have been invented independently multiple times.

Pronunciation

- IPA(key): /frɛ.nɪ.mi/

Noun

frenemy (*plural* **frenemies**)

1. (*humorous*) Someone who pretends to be your friend, but is really your enemy. [quotations ▼]
2. (*humorous*) A fair-weather friend who is also a rival.

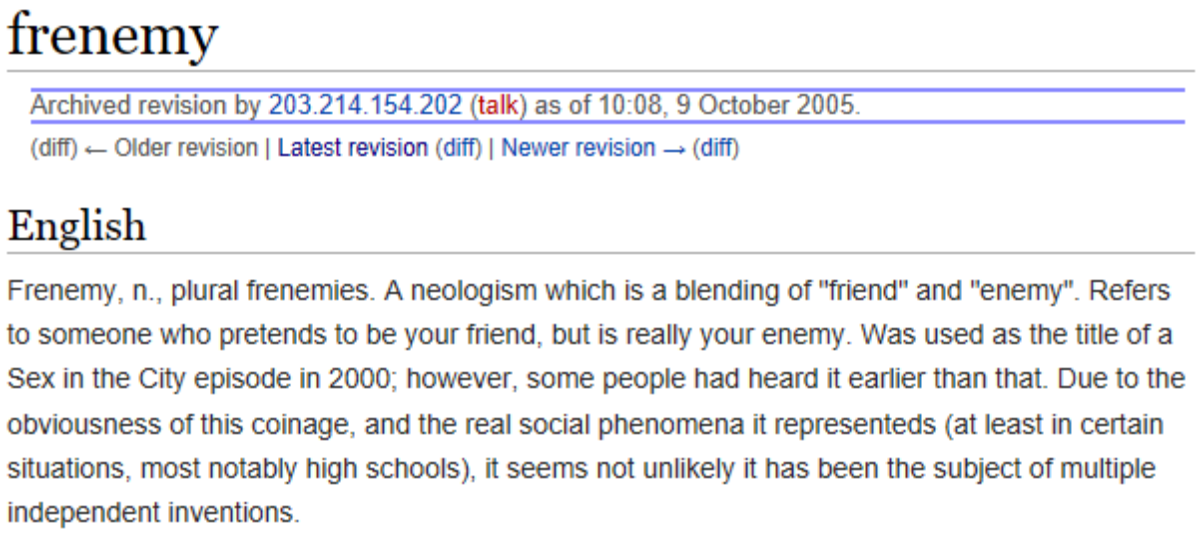
Synonyms

- betrayer
- double-crosser
- traitor
- palhole

Translations

enemy pretending to be a friend [show ▼]

Figure 5.28: *Wiktionary* entry for 'frenemy' as at 31 August 2014



frenemy

Archived revision by 203.214.154.202 (talk) as of 10:08, 9 October 2005.

(diff) ← Older revision | Latest revision (diff) | Newer revision → (diff)

English

Frenemy, n., plural frenemies. A neologism which is a blending of "friend" and "enemy". Refers to someone who pretends to be your friend, but is really your enemy. Was used as the title of a Sex in the City episode in 2000; however, some people had heard it earlier than that. Due to the obviousness of this coinage, and the real social phenomena it representeds (at least in certain situations, most notably high schools), it seems not unlikely it has been the subject of multiple independent inventions.

Figure 5.29: Original *Wiktionary* entry for 'frenemy', dated 9 October 2005

The original entry for 'frenemy' was completely unstructured, comprising a single paragraph of text containing all of the information necessary to a simple entry, but with none of the required formatting. Thus the entry included headword, word class, grammatical information (plural form), definition and word formation information. However the quality of the entry was poor because all of this was mixed together. By the time we reach 2014, a proper entry had been created (indeed it had been created in March 2006, although it was later expanded, for example by the addition of synonyms).

While it is not possible to track changes in entries in any of the other dictionaries in this way, as mentioned earlier in this section, we can see some additions when we compare screenshots taken in August 2014 with current entries, for example for 'e-tailer' in the *Oxford English Dictionary* (*OED*). Figure 5.30 shows the earlier entry, and Figure 5.31 the current one.

e-tailer, *n.*

Text size: A A

View as: Outline | Full entry

Quotations: Show all | Hide all

Pronunciation: Brit. /ˈiːteɪlə/ , U.S. /ˈiˌteɪlər/

Forms: 19– **E-tailer**, 19– **e-Tailer**, 19– **e-tailer**, 19– **eTailer**, 19– **etailer**.

Etymology: < E- *comb. form* + *-tailer* (in RETAILER *n.*). Compare E-TAILING *n.*

An organization or individual selling products or services via electronic media, esp. the Internet. Cf. E-TAILING *n.*

Categories »

1995 *Discount Store News* (Nexis) 18 Sept. 56/1 (*headline*) Wal-Mart, E-tailers headline net confab.

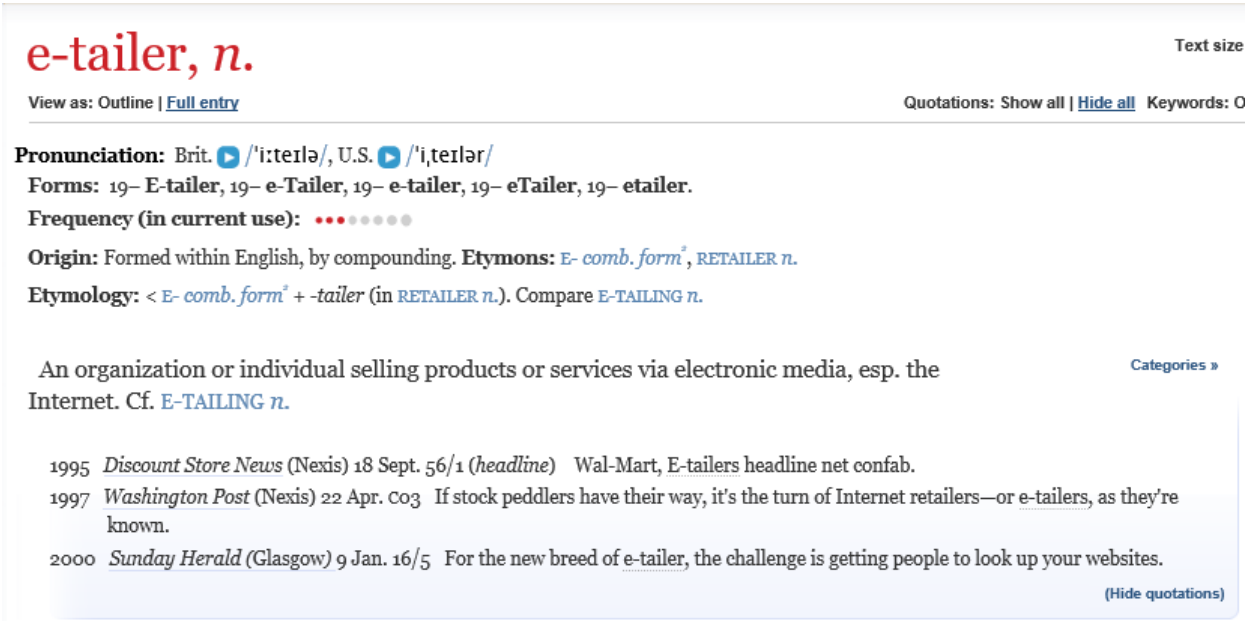
1997 *Washington Post* (Nexis) 22 Apr. C03 If stock peddlers have their way, it's the turn of Internet retailers—or e-tailers, as they're known.

2000 *Sunday Herald* (Glasgow) 9 Jan. 16/5 For the new breed of e-tailer, the challenge is getting people to look up your websites.

(Hide quotations)

Back to top

Figure 5.30: *OED* entry for 'e-tailer' 2014



e-tailer, *n.*

Text size

View as: Outline | Full entry

Quotations: Show all | Hide all Keywords: O

Pronunciation: Brit. /ˈiːteɪlə/, U.S. /ˈiˌteɪlər/

Forms: 19– **E-tailer**, 19– **e-Tailer**, 19– **e-tailer**, 19– **eTailer**, 19– **etailer**.

Frequency (in current use):

Origin: Formed within English, by compounding. **Etymons:** E- *comb. form*, RETAILER *n.*

Etymology: < E- *comb. form* + *-tailer* (in RETAILER *n.*). Compare E-TAILING *n.*

An organization or individual selling products or services via electronic media, esp. the Internet. Cf. E-TAILING *n.*

Categories »

1995 *Discount Store News* (Nexis) 18 Sept. 56/1 (*headline*) Wal-Mart, E-tailers headline net confab.

1997 *Washington Post* (Nexis) 22 Apr. C03 If stock peddlers have their way, it's the turn of Internet retailers—or e-tailers, as they're known.

2000 *Sunday Herald* (Glasgow) 9 Jan. 16/5 For the new breed of e-tailer, the challenge is getting people to look up your websites.

(Hide quotations)

Figure 5.31: *OED* entry for 'e-tailer' November 2016

When we compare the two, we can see that the more recent entry has added sound files to the pronunciation guidance at the top of the entry, and also frequency information two lines below. The word formation information below this has also been made clearer and easier to understand, increasing the quality of this dictionary component.

It appears that such changes were made to a number of *OED* entries, suggesting a website-wide update. Both the *Oxford Dictionaries* online and *Merriam-Webster* dictionary sites have been updated since data collection ceased on this project, making direct comparisons impossible.

It is clear, then, that dictionary entries do change over time, with entries in these dictionaries for newly established neologisms being relatively simple in terms of the number and quality of components they include (particularly in *Wiktionary*). This is perhaps as lexicographers and editors and contributors wait to see whether the word will prove worthy of inclusion and remain in use. When they do, in *Wiktionary* more components are added, and quickly, although in expert-produced dictionaries changes generally tend to be more of a website-wide event, such as adding frequency information to all entries, as shown above.

Through the course of this discussion then, *Wiktionary* has been shown to be the most comprehensive of the dictionaries in terms of the dictionary components its entries contain. It was found to contain more, and higher quality, dictionary components than any of the expert-produced dictionaries and these were found to develop over time. The higher levels enjoyed by *Wiktionary* are in part due to the increased flexibility it experiences through not being bound by the conventions of standardised dictionary structures. *Wiktionary* contributors are free to add different components not accepted in expert-produced dictionaries, and to do so whenever they please. Thus detailed, attestational illustrative examples can be drawn from sources that would not be permitted in traditional dictionaries. The collaborative nature of *Wiktionary* means that multiple contributors can work on each entry, offering multiple viewpoints and a wealth of knowledge that can aid the creation of more detailed entries. This provides an unmatched level of responsiveness to neologisms in use in the world at large.

Immediate updating of the site (to be discussed in the following sections, where the remaining aspects of Research Question 1 will be addressed) further consolidates this claim. However it is not simply a case of *Wiktionary* having more dictionary components; it has already been shown that *Wiktionary* uses the same defining styles as expert dictionaries (this will be discussed in full in 5.3.4). Its examples are also of the same kind as those used in traditional dictionaries. Thus *Wiktionary* is able to compete with expert-produced dictionaries both within the forum of standardised dictionary structure, and outside of it.

In addition, it has been demonstrated that there is potentially a great deal of information in the *Oxford English Corpus* (*OEC*) which could be used to expand the number of neologism entries in 'corpus-based' and even 'corpus-informed' dictionaries. However this information is not being used, to the extent that words which could be entered into the dictionary, such as 'predatory lending' are not. It is unclear why this is the case.

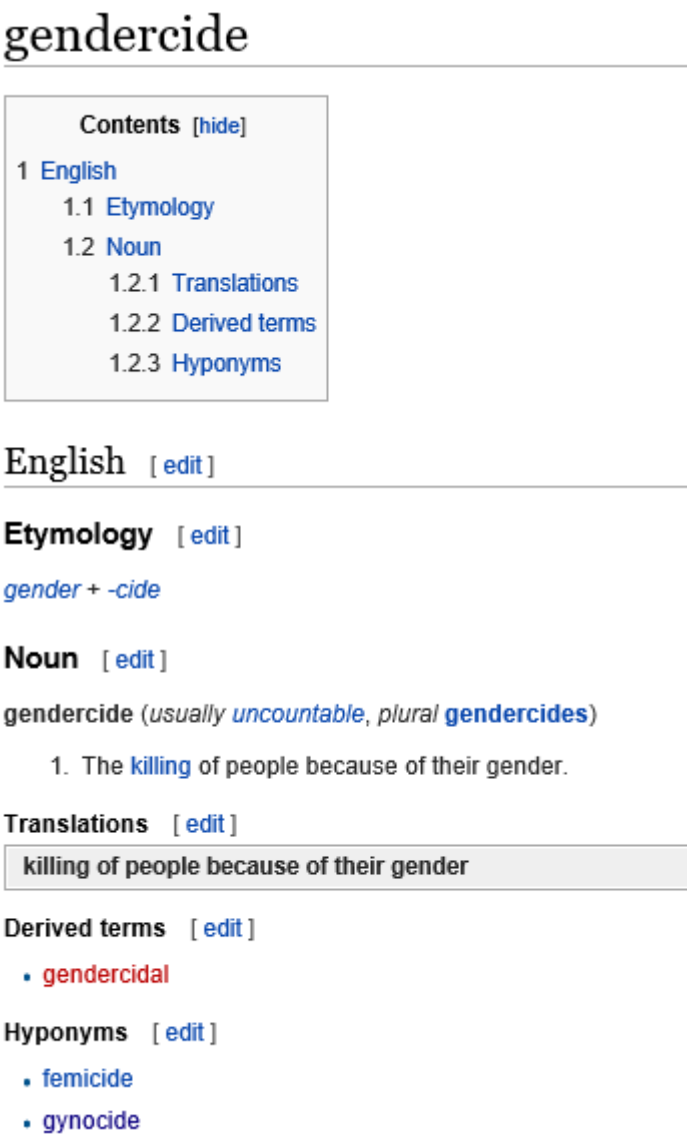
5.3.3 Transparency in Wiktionary

Two key dictionary components which *Wiktionary* alone possesses, and which provide it with an unmatched level of transparency in this study, are its Revision Histories and Discussion Forums. In this subsection I outline how these two mechanisms of the collaborative nature of *Wiktionary* make it more responsive to neologisms than the expert-produced dictionaries by allowing any contributor to make changes to any word at any time, each of which is published immediately. The transparency of the site means that other contributors can immediately see these changes and, if necessary respond to them, which in turn increases the responsiveness of the site. As mentioned above, this further consolidates the knowledge gained during this study, thus answering Research Question 1:

> *What can be learnt from this study about* Wiktionary's *responsiveness to neologisms and the level of detail and quality of definitions in its new word entries, when compared with expert-produced dictionaries?*

Within *Wiktionary* a less formal approach to style is achieved through an overarching atmosphere of consensus amongst contributors, and the provision of guidelines on how entries should look, rather than hard and fast rules (Meyer and Gurevych 2012: 273-4).

As a consequence, we see a marked variation in the dictionary components used in each entry, and sometimes significant differences in how *Wiktionary* entries actually look (see for example Figures 5.32 and 5.33).



gendercide

Contents [hide]

1 English
- 1.1 Etymology
- 1.2 Noun
  - 1.2.1 Translations
  - 1.2.2 Derived terms
  - 1.2.3 Hyponyms

English [edit]

Etymology [edit]

gender + *-cide*

Noun [edit]

gendercide (*usually uncountable*, *plural* **gendercides**)

1. The killing of people because of their gender.

Translations [edit]

killing of people because of their gender

Derived terms [edit]

- gendercidal

Hyponyms [edit]

- femicide
- gynocide

Figure 5.32: *Wiktionary* entry for 'gendercide'

bankster

Contents [hide]

1 English
- 1.1 Etymology
- 1.2 Pronunciation
- 1.3 Noun
  - 1.3.1 Translations
- 1.4 References

English [edit]

Etymology [edit]

Blend of *banker* + *gangster*

Judge Ferdinand Pecora has been credited with coining the term Bankster. In June 1933, his image appeared on the cover of Time magazine, seated at a US Senate table, a cigar in his mouth. Pecora's hearings were said to have coined a new phrase, "banksters" for the finance "gangsters." However, the word, with the same meaning, had appeared in the U.S. press at least a year and a half previous to that.

The term was later used by Léon Degrelle, Belgian fascist politician and journalist, in 1937 as a pejorative term for high financiers.

Pronunciation [edit]

- (*UK*) IPA(key): /ˈbæŋkstə/

Noun [edit]

bankster (*plural* **banksters**)

1. (*informal*, *derogatory*) A banker who is seen as criminally irresponsible, or as extorting bailout money from the taxpayers. [quotations ▼]

Translations [edit]

Translations [show ▼]

References [edit]

- The man who busted the banksters, *Smithsonian Magazine*

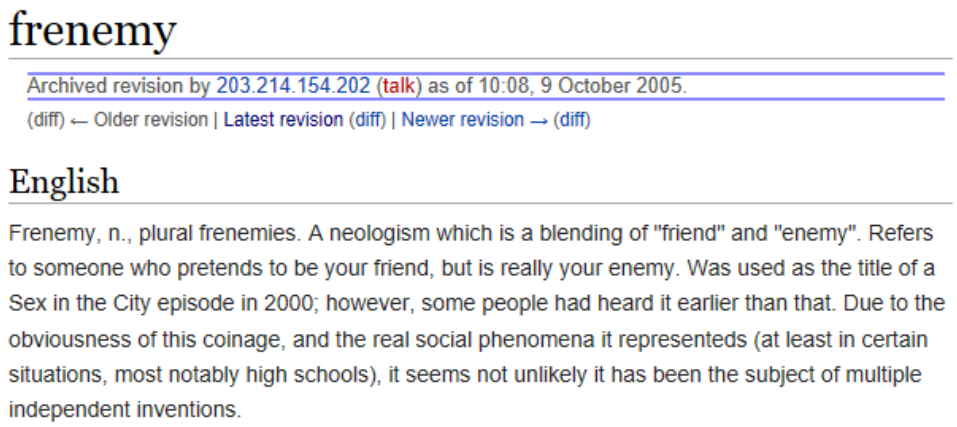
Figure 5.33: *Wiktionary* entry for 'bankster'

As we can see, the two entries look quite different and are an example of the less formal attitude to style that is part of *Wiktionary*'s responsiveness to new words. For example, the 'gendercide' entry in Figure 5.32 contains derived terms, hyponyms (unnecessary given its genus-differentiae definition style (see 5.3.4)) and a translation section in the middle of the page. Figure 5.33's 'bankster' entry looks quite different, with its long etymology section at the top and translations at the bottom. Strictly speaking, the 'gendercide' entry is in the wrong order, since translations should appear

below hyponyms and derived terms[116], for example. However the relaxed attitude to layout which contributes to *Wiktionary*'s responsiveness and detailed nature (by allowing any other contributor to come along and change the entry) means that as yet no-one has done anything about this.

Further advice on editing existing entries[117] is provided by *Wiktionary*, and 'sandboxes' are available for users to practice their work[118] without risking damaging existing entries or publishing unsuitable drafts.

The need for this was well demonstrated in the case of the entry for 'frenemy' which, as discussed in 5.3.2, was originally published with no formatting at all. It was entered by an unknown, unregistered user (registered user's names show on entries they have amended), as in Figure 5.34.


frenemy

Archived revision by 203.214.154.202 (talk) as of 10:08, 9 October 2005.

(diff) ← Older revision | Latest revision (diff) | Newer revision → (diff)

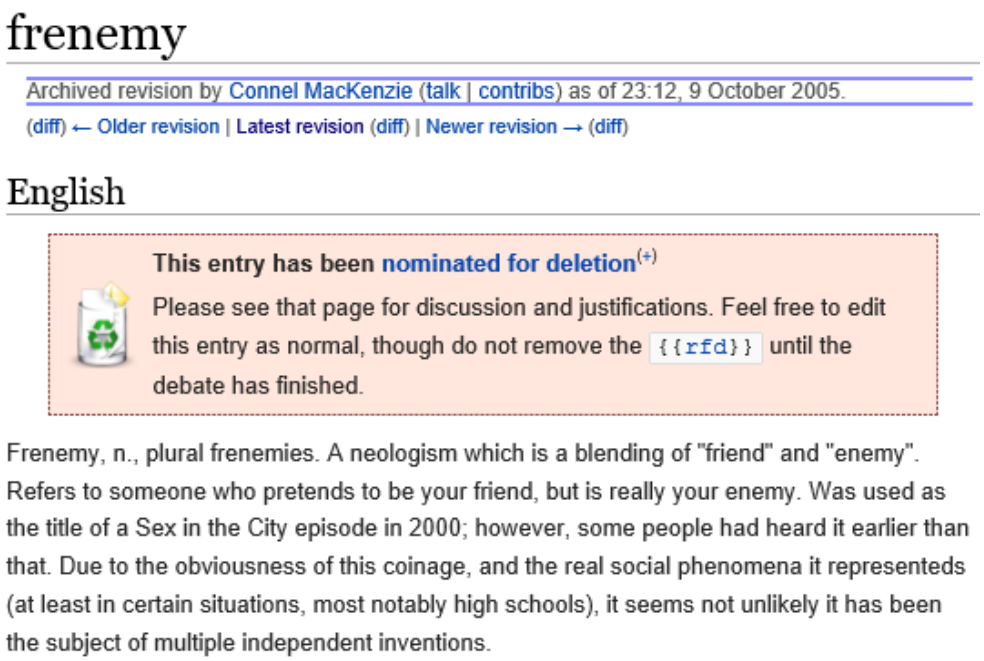
English

Frenemy, n., plural frenemies. A neologism which is a blending of "friend" and "enemy". Refers to someone who pretends to be your friend, but is really your enemy. Was used as the title of a Sex in the City episode in 2000; however, some people had heard it earlier than that. Due to the obviousness of this coinage, and the real social phenomena it representeds (at least in certain situations, most notably high schools), it seems not unlikely it has been the subject of multiple independent inventions.

Figure 5.34: Original *Wiktionary* entry for 'frenemy', in non-standard format

Later that same day, the entry was nominated for deletion, something which can only be done through a request for a consensus decision to delete (see Figure 5.35)[119]

[116] See https://en.wiktionary.org/wiki/Wiktionary:Entry_layout#
[117] https://en.Wiktionary.org/wiki/Help:How_to_edit_a_page
[118] See for example https://simple.Wiktionary.org/wiki/Wiktionary:Sandbox
[119] https://en.Wiktionary.org/wiki/Wiktionary:Requests_for_deletion



frenemy

Archived revision by Connel MacKenzie (talk | contribs) as of 23:12, 9 October 2005.
(diff) ← Older revision | Latest revision (diff) | Newer revision → (diff)

English

This entry has been nominated for deletion(+)
Please see that page for discussion and justifications. Feel free to edit this entry as normal, though do not remove the {{rfd}} until the debate has finished.

Frenemy, n., plural frenemies. A neologism which is a blending of "friend" and "enemy". Refers to someone who pretends to be your friend, but is really your enemy. Was used as the title of a Sex in the City episode in 2000; however, some people had heard it earlier than that. Due to the obviousness of this coinage, and the real social phenomena it representeds (at least in certain situations, most notably high schools), it seems not unlikely it has been the subject of multiple independent inventions.

Figure 5.35: *Wiktionary* original entry for 'frenemy' put forward for deletion

Rather than consensus being reached to delete 'frenemy', the next day the entry was moved to a 'request for verification', indicating that a user did not believe it met the criteria to become a *Wiktionary* entry (see Figure 5.36).

frenemy

Archived revision by Muke (talk | contribs) as of 01:04, 10 October 2005.
(diff) ← Older revision | Latest revision (diff) | Newer revision → (diff)

English

A user has added this entry to requests for verification(+)
If it cannot be verified that this term meets our attestation criteria, then it will be deleted. Feel free to edit this entry as normal, but do not remove {{rfv}} until the request has been resolved.

Frenemy, n., plural frenemies. A neologism which is a blending of "friend" and "enemy". Refers to someone who pretends to be your friend, but is really your enemy. Was used as the title of a Sex in the City episode in 2000; however, some people had heard it earlier than that. Due to the obviousness of this coinage, and the real social phenomena it representeds (at least in certain situations, most notably high schools), it seems not unlikely it has been the subject of multiple independent inventions.

Figure 5.36: *Wiktionary* entry for 'frenemy' with a request for verification

A few hours later, four citations had been found for the word, and the entry had been converted into the preferred format, as shown in Figure 5.37.

frenemy

Archived revision by Muke (talk | contribs) as of 06:33, 10 October 2005.
(diff) ← Older revision | Latest revision (diff) | Newer revision → (diff)

Contents [hide]
1 English
1.1 Etymology
1.2 Noun
1.3 Quotations

English

Etymology

A blend of "friend" and "enemy". Likely to have been invented independently multiple times.

Noun

frenemy *(plural:* frenemies.)

1. Someone who pretends to be your friend, but is really your enemy.

Quotations

- **2000**: **frenemies** —*Sex and the City,* season 3 episode 16, first aired October 1. [title]
- **2001**: In France the Seine has all the advantages of Northernness (a quality underrated by our Gallic **frenemy**) but it is too fatally interested in Paris [...] —John Lanchester, *The Debt to Pleasure.* [1]
- **2004:** You know when you dump a guy, only to discover years later that he's evolved into the perfect boyfriend—for the high-school **frenemy** who convinced you to dump him in the first place...? —*The Ex-Factor,* Andrea Semple. [back cover] [2]
- **2005:** So why did we break up? Enter Blaize St. John, **frenemy** extraordinaire. She came, she saw, she stole my boyfriend. —*Single Girl's Guide to Murder,* Joanne Meyer. [back cover] [3]

Figure 5.37: *Wiktionary* entry for 'frenemy' including quotations

From this sequence of events we can see the procedure by which new words not only enter *Wiktionary*, but their inclusion can be questioned and validated through consensus and debate. While no actual Discussion page (see below) was begun on the topic, consensus was reached through collaboration between independent contributors. The entire process took place in less than 24 hours, demonstrating the speed at which *Wiktionary* responds to neologisms. As soon as a word meets the inclusion criteria an entry can be created in the dictionary, published and ready for use. In any of the expert-produced dictionaries under discussion here, 'frenemy' would have been added to a list of words ready for inclusion at the next update. In the Oxford suite of dictionaries, for electronic versions, this would have taken at most three months, as

demonstrated by the 'Recent Updates to *OED*' menu (which, confusingly, also serves *Oxford Dictionaries* online), shown in Figure 5.38. (The period between publication of the current printed edition of *ODE* and the previous one was seven years, and the future of Oxford dictionaries in printed form is uncertain[120].)

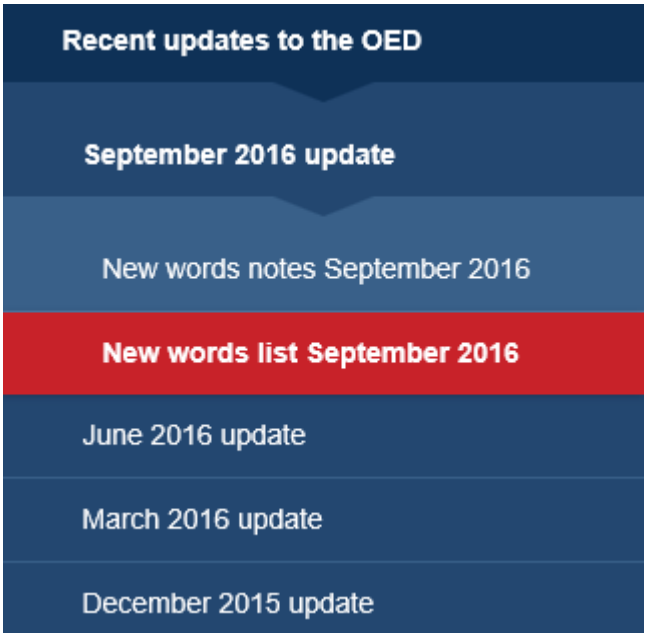


Figure 5.38: 'Recent Updates' menu for *OED* and other Oxford electronic dictionaries[121]

In *Merriam-Webster*, the delay between choosing words for inclusion and adding them to the dictionary appears to be an entire year, based upon the date information on the *MW* website for 2015 and 2016 (we can tell these additions are for the electronic version based on the 'sound file' symbols on for example 'emoji'[122]).

Another element of transparency that we find with *Wiktionary*, but which is absent from all of the other dictionaries, is the ability to see who is making changes to entries, as well as who is entering neologisms. If we examine the Revision History for 'waterboarding', we see that it entered *Wiktionary* as a noun in January 2007, as shown in Figure 5.39.

[120] See for example http://www.telegraph.co.uk/culture/culturenews/10777079/RIP-for-OED-as-worlds-finest-dictionary-goes-out-of-print.html
[121] http://public.OED.com/the-*OED*-today/recent-updates-to-the-OED/september-2016-update/new-words-list-september-2016/
[122] See http://unabridged.Merriam-Webster.com/blog/2015/05/a-growth-spurt/

waterboarding

Archived revision by Language Lover (talk | contribs) as of 14:54, 27 January 2007.
(diff) ← Older revision | Latest revision (diff) | Newer revision → (diff)

English

Noun

Wikipedia has an article on: waterboarding

waterboarding (*plural* **waterboardings**)

1. A type of torture employed by the U.S. and its allies, where the victim is immobalized, has rags placed over their face, and has water poured thereinto, which creates the sensation of drowning.

Figure 5.39: *Wiktionary* entry for 'waterboarding' as a noun

On 26 April 2009, 'torture' was changed to 'harsg interrogation' [sic]. The spelling error was corrected by another user, however for the rest of the day several users went back and forth between 'torture' and 'harsh interrogation', until finally the entry was left as it had begun. It remained unchanged until 2 December 2012, when 'torture' became 'torture technique', which it remains to this day, although the rest of the definition has since been expanded.

Had an error been introduced into any of the other dictionaries here, however, it would have been months before it could be corrected, and this again, is an example of the levels of responsiveness which *Wiktionary* can achieve and expert-produced dictionaries cannot (although it should be noted that traditional dictionaries are proof-read by professionals, who would hopefully spot and correct such an error). It is also worth bearing in mind that had the changes to the 'waterboarding' entry been published to a 'beta' site first (available only to registered users), it might have been possible to avoid the publication of the spelling error to the general population at all. In my personal experience managing commercial websites, a 'beta' site is always used as a tool for proof-reading and checking changes before releasing them to the general public. Thus the immediacy of response enjoyed by *Wiktionary* is not without its occasional negative consequences.

Returning to Research Question 1, then, we can say that this study has demonstrated just how much more responsive to neologisms *Wiktionary* is than the expert-produced dictionaries, but also the occasional negative impact that can arise from this.

Research Question 1 – *What can be learnt from this study about* Wiktionary's *responsiveness to neologisms and the level of detail and quality of definitions in its new word entries, when compared with expert-produced dictionaries?*

Through these records of changes to entries, it is possible to begin to identify which contributors' opinions and amendments can be relied upon, and which cannot. This is useful, since although there is the chance for registered users to provide profile information about their activities, both within and outside of *Wiktionary*, many prefer privacy, and hence little is known about them (see below). Crucially, it is rare to know what, if any, lexicographical knowledge or experience a contributor may have. Regular use of *Wiktionary* therefore allows a sense of the expertise (or otherwise) of certain contributors to become apparent, and at the same time we see a kind of community of participants developing, similar to that claimed to be present among '*Wikipedians*' (Bryant, Forte and Bruckman 2005).

This is the case when we examine the contributors responsible for first entering the neologisms in this study; 'SemperBlotto', for example, began the entries for 'e-tailer', 'e-waste', 'hubristic', 'upskill', 'wellderly', 'diabesity' and 'sovereign debt'. The first of these was in 2006, the last in 2011. SemperBlotto's real name is Jeff Knaggs, and as well as entering neologisms, he is also involved in developing the Latin, French, German and Italian indices in *Wiktionary*[123]. He is also an administrator of the site (administrators are contributors who have been nominated and elected to the post, which gives them additional access, for example to block articles/users or to delete pages from the site (Meyer and Gurevych 2012: 271)). He is involved in hundreds of '*WikiMedia*' projects[124], as well as also being a contributor to *Wikipedia*. While clearly a language enthusiast, there is no indication that Knaggs has any formal training in linguistics, and an internet search for him returns no valid results. However the overall impression one gets, after time spent browsing pages in which he has been involved, is that his contributions are generally of high quality.

[123] See https://en.Wiktionary.org/wiki/User:SemperBlotto
[124] See https://meta.wikimedia.org/wiki/Special:CentralAuth/SemperBlotto

Like several prolific *Wiktionary* contributors, SemperBlotto also uses a 'bot', 'a piece of software designed to complete a minor but repetitive task automatically or on command, especially when operating with the appearance of a (human) user profile or account'[125]. Knaggs' bot trawls through *Wiktionary* adding verb translations and some noun and adjective forms in several languages including Italian[126]. Most of the 'bots' I have found in *Wiktionary* appear to serve this sort of function, adding translation information in a variety of different languages. When viewing the Revision History for a word, one will often find that many of the entries are by bots rather than people, as shown in Figure 5.40, from the entry for 'cyberbullying'.

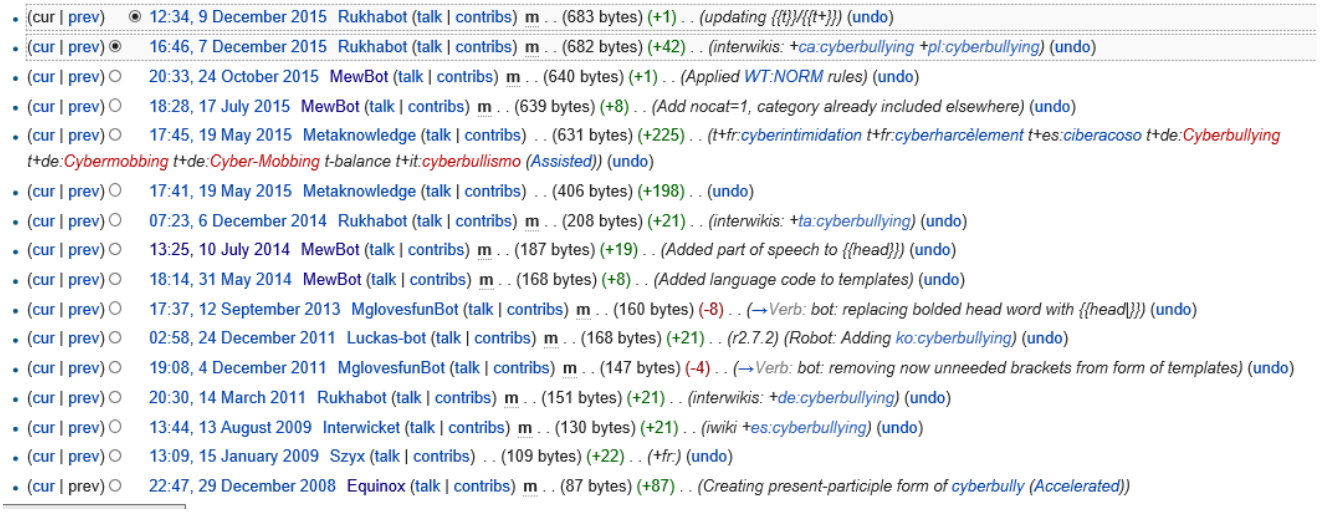

(cur | prev) ◉ 12:34, 9 December 2015 Rukhabot (talk | contribs) m . . (683 bytes) (+1) . . *(updating {{t}}/{{t+}})* (undo)
(cur | prev) ◉ 16:46, 7 December 2015 Rukhabot (talk | contribs) m . . (682 bytes) (+42) . . *(interwikis: +ca:cyberbullying +pl:cyberbullying)* (undo)
(cur | prev) ○ 20:33, 24 October 2015 MewBot (talk | contribs) m . . (640 bytes) (+1) . . *(Applied WT:NORM rules)* (undo)
(cur | prev) ○ 18:28, 17 July 2015 MewBot (talk | contribs) m . . (639 bytes) (+8) . . *(Add nocat=1, category already included elsewhere)* (undo)
(cur | prev) ○ 17:45, 19 May 2015 Metaknowledge (talk | contribs) . . (631 bytes) (+225) . . *(t+fr:cyberintimidation t+fr:cyberharcèlement t+es:ciberacoso t+de:Cyberbullying t+de:Cybermobbing t+de:Cyber-Mobbing t-balance t+it:cyberbullismo (Assisted))* (undo)
(cur | prev) ○ 17:41, 19 May 2015 Metaknowledge (talk | contribs) . . (406 bytes) (+198) . . (undo)
(cur | prev) ○ 07:23, 6 December 2014 Rukhabot (talk | contribs) m . . (208 bytes) (+21) . . *(interwikis: +ta:cyberbullying)* (undo)
(cur | prev) ○ 13:25, 10 July 2014 MewBot (talk | contribs) m . . (187 bytes) (+19) . . *(Added part of speech to {{head}})* (undo)
(cur | prev) ○ 18:14, 31 May 2014 MewBot (talk | contribs) m . . (168 bytes) (+8) . . *(Added language code to templates)* (undo)
(cur | prev) ○ 17:37, 12 September 2013 MglovesfunBot (talk | contribs) m . . (160 bytes) (-8) . . *(→Verb: bot: replacing bolded head word with {{head|}})* (undo)
(cur | prev) ○ 02:58, 24 December 2011 Luckas-bot (talk | contribs) m . . (168 bytes) (+21) . . *(r2.7.2) (Robot: Adding ko:cyberbullying)* (undo)
(cur | prev) ○ 19:08, 4 December 2011 MglovesfunBot (talk | contribs) m . . (147 bytes) (-4) . . *(→Verb: bot: removing now unneeded brackets from form of templates)* (undo)
(cur | prev) ○ 20:30, 14 March 2011 Rukhabot (talk | contribs) m . . (151 bytes) (+21) . . *(interwikis: +de:cyberbullying)* (undo)
(cur | prev) ○ 13:44, 13 August 2009 Interwicket (talk | contribs) m . . (130 bytes) (+21) . . *(iwiki +es:cyberbullying)* (undo)
(cur | prev) ○ 13:09, 15 January 2009 Szyx (talk | contribs) . . (109 bytes) (+22) . . *(+fr:)* (undo)
(cur | prev) ○ 22:47, 29 December 2008 Equinox (talk | contribs) m . . (87 bytes) (+87) . . *(Creating present-participle form of cyberbully (Accelerated))*

Figure 5.40 Excerpt from Revision History for 'cyberbullying'

Out of the 16 changes shown on the Revision History for 'cyberbullying' in Figure 5.40, 11 are by 'bots', mostly 'Rukhabot', which deals with wiki links and templates[127] and 'Mewbot', which adds verb translations in Catalan, Esperanto, Finnish and Dutch[128].

'Mewbot' is run by someone calling themselves 'C*ODE*Cat' (no real name given). Like SemperBlotto, s/he is clearly a language enthusiast and is a *Wiktionary* administrator, but again, there is no evidence of any formal linguistic training[129]. Ruakh is a prolific *Wiktionary* contributor but provides no profile information at all.

[125] https://en.Wiktionary.org/wiki/bot
[126] https://en.Wiktionary.org/wiki/User:SemperBlotto
[127] https://en.Wiktionary.org/wiki/User:Rukhabot
[128] https://en.Wiktionary.org/wiki/User:MewBot
[129] See https://en.Wiktionary.org/wiki/User:CODECat

It would seem that the bots adding translation information are essentially attempting to turn *Wiktionary* into a multilingual dictionary, since there are already hundreds of monolingual versions of *Wiktionary* in different languages (Meyer and Gurevych 2012: 262-267). While they add components to neologism entries, it is difficult to assess their quality. It is clear that the information provided is limited, however, since in many cases all that is added is the headword itself, with no translation of any of the other information from the entry, for example the definition. See for example the entry for 'promissory note' in which none of the translations seems long enough to cover all of the entry information[130].

A more useful bot might perhaps be one which periodically searched new entries and identified key lexicographical components which had still yet to be added, for example 'wellderly' which still has no pronunciation information, sound file or subject label six years after entering the dictionary.

Thus through Revision Histories it is possible to identify the most active and most reliable *Wiktionary* contributors, many of whom tend to also be administrators (although administrator status is not highly publicised, I suspect as part of the non-hierarchical organisation of the site (Wiktionary 2016d)). It is also possible to cross reference other pages or discussions in which these contributors have been active.

The other main elements of the collaborative culture of *Wiktionary* are the Discussion Forums. All DDEB2 and DDEB3 neologisms except 'acedia', 'greenwashing' (which is not included in *Wiktionary*) and 'promissory note' generate search results through the Tea Room archive, however most of these results are simply references to the word in question. For example searching for 'bankster' generates a result for the 'banksterism'[131] entry, a word which we would really expect to see as a related term in the main entry for 'bankster'[132], since it is an example of the kind of lexical creativity discussed by Fischer (1998), Renouf (2007) and Moon (2008). Yet it is not present. Similarly 'hyperlocal' brings up 'hyperlocalism'[133], another example of lexical creativity,

[130] https://en.Wiktionary.org/w/index.php?title=promissory_note&oldid=33637271
[131] https://en.wiktionary.org/wiki/banksterism
[132] https://en.wiktionary.org/wiki/bankster
[133] https://en.wiktionary.org/wiki/hyperlocalism

yet also not presented as a related term. This is one of the difficulties of a collaborative dictionary like *Wiktionary*; because there is no editorial oversight, connections like this can be missed.

None of the neologisms under study have their own discussion threads in the Tea Room, although 'hubristic' and 'promissory note' are both mentioned in other *Wiktionary* entries' threads. 'Frenemy' brings up a Talk page[134] where an unregistered user in May 2015 suggests adding to the entry that the earliest usage of the term was either in print or television/film, and 'upskill' carries a Talk suggestion[135] about creating '{{horrible word}} templates' that indicates that the (unregistered) user has not really understood the purpose of the Talk function. Neither suggestion was ever taken up, and in the case of 'frenemy', the entry instead states that it is 'likely to have been invented independently multiple times'. Several quotations are included, although there is no indication of whether any of these is believed to be the 'first'.

Such quotations (and references) both serve as examples of the headword (see 5.3.2) and provide attestational evidence of the use of the neologism which likely contributed to its acceptance into the dictionary, as shown in Figure 5.41.

[134] https://en.wiktionary.org/wiki//Talk:frenemy
[135] https://en.wiktionary.org/wiki//Talk:upskill

superphone

English [edit]

Etymology [edit]

super- + *phone*

Noun [edit]

superphone (*plural* **superphones**)

1. (*informal*) A remarkably advanced telephone. [quotations ▲]
   - **1978**, *Popular Mechanics* (volume 149, number 4, April 1978)
     Other single-unit **superphones** are Figure-phone, which packs a calculator, clock and automatic redialer for about $270...
   - **1981**, *Science Digest* (volume 89)
     A **"superphone"** for the deaf and deaf-blind, created by Ultratech of Madison, Wisconsin, can convert a typed message into sound...
   - **1989**, Austin H. Kiplinger, Knight A. Kiplinger, *America in the global '90s: the shape of the future*
     The same **superphones** will monitor the security system in your home...
   - **2005**, William Charles Mann, *Smart technology for aging, disability, and independence*
     Not everyone owns a **"superphone"** with internet access like the Sony Ericsson P900...
   - **2008**, *Mac Life* (volume 2, number 4, April 2008)
     Standing around pushing buttons on a **superphone** creates quite an appetite. So I stopped at my favorite taqueria...
   - **2009**, Alastair Sweeny, *BlackBerry Planet*
     The device was a truly smart phone—some even said a **superphone**. Jobs argued that it had enough security to satisfy all but the most paranoid IT manager.

Figure 5.41: *Wiktionary* entry for 'superphone'

As we can see, the quotations for the use of 'superphone' date back to 1978, although I suspect the 'superphone' referred to in that first quotation will have been a very different product to the 'superphones' of today. These attestation sources can come from any *Wiktionary* contributor wishing to add to the entry, and through this we again see the collaborative nature of *Wiktionary*, coupled with its less formal approach to inclusion and style. *Wiktionary*'s Revision Histories and Discussion Forums (examined during the course of this study in order to answer Research Question 1) help to make it more responsive to neologisms than any of the other dictionaries can hope to be, in part, as expected, due to the speed of updating of the site. By enabling new words to be

on the site and in use within hours of first meeting the selection criteria, and to have discussion and amendments proceeding even while the entry is in place, are advantages which only a collaborative dictionary could enjoy.

The truly 'collaborative', non-hierarchical approach to the running of *Wiktionary* is very different to the traditional 'top down' organisation of expert-produced dictionaries, which often have a team of Editors working beneath a Managing Editor, an Editor-in-Chief and a Publisher. We may presume that funding and the logistical issues inherent in being part of a large publishing organisation, as well as much stricter inclusion criteria, prevent the expert-produced dictionaries from being updated more than three or four times a year. On the other hand the collaborative nature of *Wiktionary* means that it can respond to a neologism literally on the day that it is presented for inclusion.

5.3.4 Neologism Definitions: Comparisons Between Different Dictionaries

In this section I present and discuss findings from comparisons made between neologism definitions from different dictionaries and different dictionary types. These findings, and the element of Research Question 1 dealing with definitions, are addressed here, slightly separately from the other dictionary components, due to the importance of good quality definitions in the content of any dictionary entry.

The subjective assessments of dictionary definitions comprised deciding the degree to which they matched one another, in spirit if not in exact language, the type of definition used (see 3.4.3) and in particular how comprehensive each definition was in comparison to that of *Wiktionary*. Shared elements or concepts in definitions were noted, and in particular original *Wiktionary* definitions were compared with those present as at 31 August 2014, in order to assess whether any significant changes had taken place during their years of inclusion on the website. The definitions in the five dictionaries were compared with one another, and also with that provided by the *NeoCrawler* program, from which the list of neologisms used in the study was originally derived. It had already been determined, however, that some *NeoCrawler* definitions were poorly written, overly complicated or too simplistic.

When I first began to examine the definitions I found that in most cases the meaning ascribed to the neologisms was much the same across the different dictionary titles and types. At the same time, all of the dictionaries utilised the same defining styles, most of the neologisms being defined according to the classical 'genus-differentiae' model (Atkins and Rundell 2008: 414). For example 'waterboarding' is consistently described as a type of torture that simulates drowning, as shown in Figures 5.42 – 5.46.

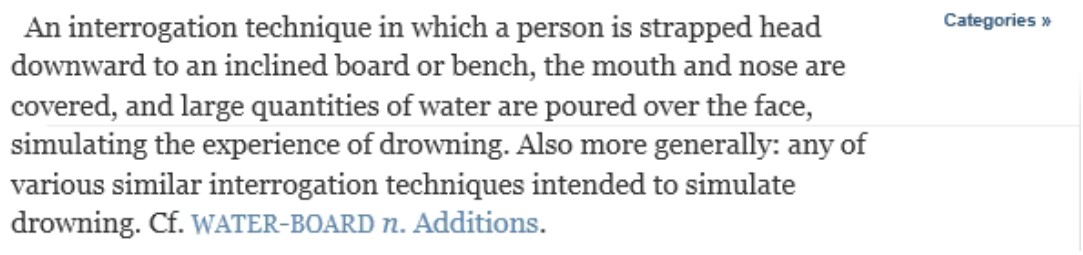


Figure 5.42: *OED* definition for 'waterboarding'

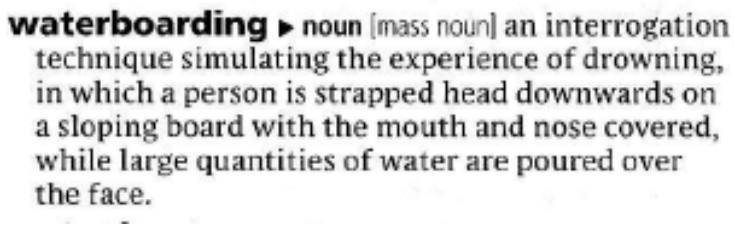


Figure 5.43: *ODE* definition for 'waterboarding'

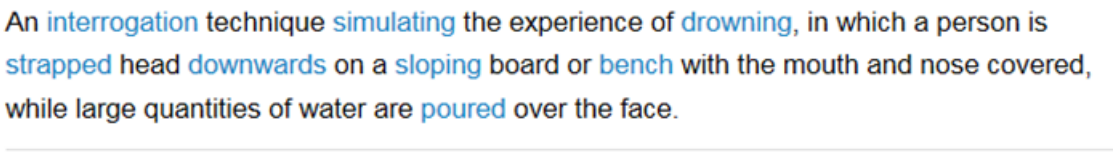


Figure 5.44: *ODO* definition for 'waterboarding'

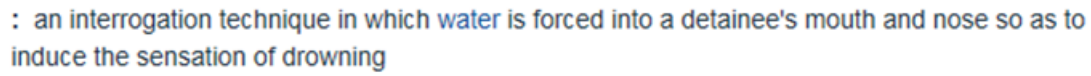


Figure 5.45: *MW* definition for 'waterboarding'

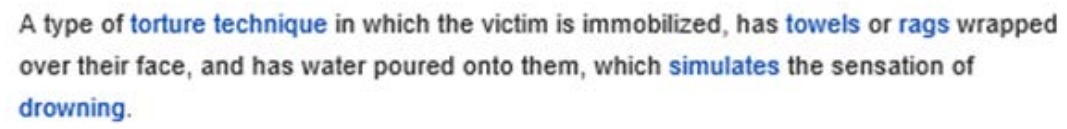


Figure 5.46: *Wiktionary* definition for 'waterboarding'

Although in some cases longer than the standard genus definition, all of the dictionaries chose the same defining strategy (as did the *NeoCrawler*: 'A torture method of putting a cloth over their face and pouring water over it to make them believe they are drowning' (*NeoCrawler* list, Ludwig-Maximilians Universität n.d.)). The only other defining style found amongst these neologism entries was defining by synonym, using a word/phrase that means the same thing (Atkins and Rundell 2008: 420-1). Thus 'tenebrous' is consistently defined as gloomy, dark, murky or obscure. Yet as Atkins and Rundell point out, no two words are truly the same (Ibid), and synonyms do not actually explain the meaning of the word. Definitions by synonym are considered to be successful only when the two terms are semantically 'identical', and this is rare outside of technical contexts (Ibid: 421; Svensén 2009: 215). None of the neologisms here use any of the other defining strategies discussed in 3.4.3.

It is interesting that contributors to collaborative dictionaries choose to define words in the same ways as trained lexicographers. There is no guidance on *Wiktionary* about how a definition should be worded, and hence it seems that contributors automatically follow the style they are familiar with from other dictionaries. Although they make no reference to defining styles, Meyer and Gurevych conclude that *Wiktionary* does make a credible rival for expert-produced dictionaries (2012: 291) and Penta agrees that 'cyber-lexicons are on par with the *OED* in handling semantic information' (2011: 10). While comparing *Wiktionary* definitions with those of expert-produced dictionaries then, we can know that we are indeed comparing like-with-like.

The one dictionary entry in which the definition differs significantly from those in all of the other dictionaries, is for one of the 'reincarnated' words (see 5.4.1.2), 'hubristic'. In the *Oxford English Dictionary* (*OED*) this is defined as 'insolent, contemptuous', whereas all the other dictionaries refer to excessive pride, arrogance or self-confidence. This is

particularly surprising since the *OED* definition for 'hubris' also refers to excessive pride or arrogance. Why the *OED* provides such a different definition for 'hubristic' is unclear. There is also not a single example of 'hubristic' being used by newspapers in accordance with the *OED*'s definition. All of the newspapers, regardless of article type, appear to adopt the 'excessive pride' or arrogance/self-confidence definition, see for example the concordance lines from my media tracking database in Figure 5.47.

lenders to Dubai World --the ruling al-Maktoum family's	**hubristic** /hubristic-j	business venture. One of the consequences of
money. Of course, the real culprit is RBS. Its	**hubristic** /hubristic-j	purchase of the Dutch bank ABN Amro, after the credit
Scotland. Sir Fred Goodwin's ill-conceived,	**hubristic** /hubristic-j	dash for greatness in the boom years has landed the
in the viral marketing campaign for it, including a	**hubristic** /hubristic-j	video lecture set in 2023. Janek (Idris Elba
themselves. "They are both opportunistic and	**hubristic** /hubristic-j	," he said. "When al-Qaida in Iraq first emerged,
world title. It is hoped this view does not prove	**hubristic** /hubristic-j	. Yet with enthusiasm, more than arrogance, Khan is
wrote a symphonic poem Wallenstein's Camp about the	**hubristic** /hubristic-j	count who built this palace to outshine the Holy
noughties shapes, which have come to represent the	**hubristic** /hubristic-j	building culture of the last few years, just as tellingly
These days, he jokes about the bullying - "I was a	**hubristic** /hubristic-j	little tosser" - but at the time he found his unpopularity
colour, creed or species. This is an audacious,	**hubristic** /hubristic-j	gambit, though I'm not sure the comparisons quite
ennoblement never made much sense, given that he OK'd the	**hubristic** /hubristic-j	Royal Bank of Scotland merger for which RBS boss
hysterical Antigone is collateral damage in Creon's	**hubristic** /hubristic-j	refusal to recognise that, as the Chorus concludes
40 years, was deposed — and David Cameron made that	**hubristic** /hubristic-j	visit with President Sarkozy of France to congratulate

Figure 5.47: Sketch Engine concordance lines for 'hubristic' from my media tracking database

Several of the articles shown above are business related, and it is difficult to imagine the meaning of 'hubristic' in this context as being 'insolent' or 'contemptuous'. My subjective sense is that it is *OED* that is out of step here. The most likely reason seems to stem from the original *OED* entry for 'hubristic' in 1899 (after which the word probably fell out of favour for many years, see 5.4.1.2). However the *OED* Publication History box for 'hubristic' suggests that, while the entry may not have been fully updated since then, there have been some changes[136], leading to further questions as to why the definition remains so unusual. I cannot answer, except to imagine that 'insolent, contemptuous' was one of the meanings of 'hubristic' over a century ago. I have, however, been unable to find any other reference to this meaning. This is a time, then, when it seems that the 'reading programmes' used by 'corpus-informed' dictionaries (see 3.4 and its subsections) as opposed to the actual corpora used by 'corpus-based' ones can result in misleading definitions. The problem is perhaps also a result of the 'staggered' *OED* updating process currently underway, with updates

[136] See http://www.*OED*.com/view/Entry/89082?redirectedFrom=hubristic#eid

starting from the letter M; this means that the letter H will be among the last to be updated, at a time as yet unspecified.

Aside from the more general similarities between definitions, there were a number of definitions which were almost exactly the same in all the dictionaries in the Oxford suite. These were 'e-waste', 'frenemy', 'promissory note' and 'waterboarding', all of which fell into Dictionary Date of Entry Batch 3 (DDEB3), meaning that they have been in dictionaries for some years, dating back to the beginning of the neologic life-cycle (see 3.9). *Wiktionary*'s entry for 'frenemy' is very similar to that of *Merriam-Webster* (*MW*) (to the extent that I suspect this is probably where it came from, although this is not listed as a citation). Both make the entry more personal than the Oxford dictionaries, placing the onus of 'non-friendship' on the other person, as opposed to the implied sense that the feeling is shared: *Wiktionary* sense 1: 'someone who pretends to be your friend, but is really your enemy'; *ODE*: 'a person with whom one is friendly despite a fundamental dislike or rivalry'. The use of the word 'pretends' in the *Wiktionary* definition suggests that the other person is being intentionally deceptive, whereas the 'fundamental dislike or rivalry' in *ODE* could come from either party (the issue of rivalry is covered by *Wiktionary* in sense 2).

As intimated above, *Wiktionary*, and *OED*, carry a second sense for 'frenemy', although the information is covered by a single sense in the other dictionaries. As discussed previously (5.3.2), the original *Wiktionary* entry for 'frenemy' was a piece of prose carrying several dictionary components but in a largely confusing style.

As would be expected, given that *ODO* is the electronic version of *ODE*, we see similarities in these definitions fairly frequently, although only in DDEB3, because none of the neologisms in DDEB2 were included in *ODE* (presumably being too new to yet meet the *ODE*'s inclusion criteria). Thus *ODE* and *ODO* share the same definitions for 'conurbation', 'BOGOF', 'warrantless', 'acedia' and 'e-tailer'. Perhaps the least satisfactory *ODE*, *ODO* and *OED* entry is the one for 'acedia', as it simply provides a cross-reference to an earlier term 'accidie'. *Wiktionary*, however, provides a full entry, with three senses, as shown in Figure 5.48.

Noun [edit]

acedia (*uncountable*)

1. spiritual or mental sloth.
2. apathy; a lack of care or interest; indifference
3. boredom

Figure 5.48: Three senses in the *Wiktionary* definition of 'acedia', from 2014 entry[137]

Clearly it is much more helpful to the standard user to see an immediate definition and not to have to follow cross references to related words in order to find out the meaning. By contrast, *Wiktionary*'s definition for 'e-tailer' is less comprehensive than that of any of the Oxford dictionaries. *ODO* and *ODE* provide the same definition: 'a retailer selling goods via electronic transactions on the internet'. Like *Merriam-Webster* (*MW*), *Wiktionary* fails to mention 'selling', and makes no mention of 'goods' or products of any kind (*MW* has corrected some of this since 2014; its entry now does include the term 'sells products'[138]). Instead, *Wiktionary* provides a much more generalised definition: 'A company that does business via electronic media, especially via the Internet'[139]. This is the same as the original definition entered in September 2006.

As one might expect given the collaborative nature of *Wiktionary*, several of its neologism definitions are clearer and more accessible than those found in either the 'corpus-based' or 'corpus-informed' expert-produced dictionaries (despite using the same defining styles). These are the definitions for 'promissory note' and 'warrantless' from DDEB3 and 'bankster', 'cyberchondriac', 'diabesity' and 'hyperlocal' from DDEB2. When we consider this in light of Research Question 1, along with the definitions of 'frenemy' and the issue of 'hubristic' in the *OED*, we can see that this study has shown that in many cases, as well as being clearer and more accessible than other definitions, these are also more detailed in the sense that they contain more information, or they present the information more clearly.

[137] https://en.Wiktionary.org/w/index.php?title=acedia&oldid=27157355
[138] http://www.*Merriam-Webster*.com/dictionary/e-tailer
[139] https://en.Wiktionary.org/w/index.php?title=e-tailer&oldid=26740046

When we look at ‘promissory note’, for example, the language used in the entry which is common to all three Oxford dictionaries is what we might expect in a legal or insurance document, as Figure 5.49 shows.

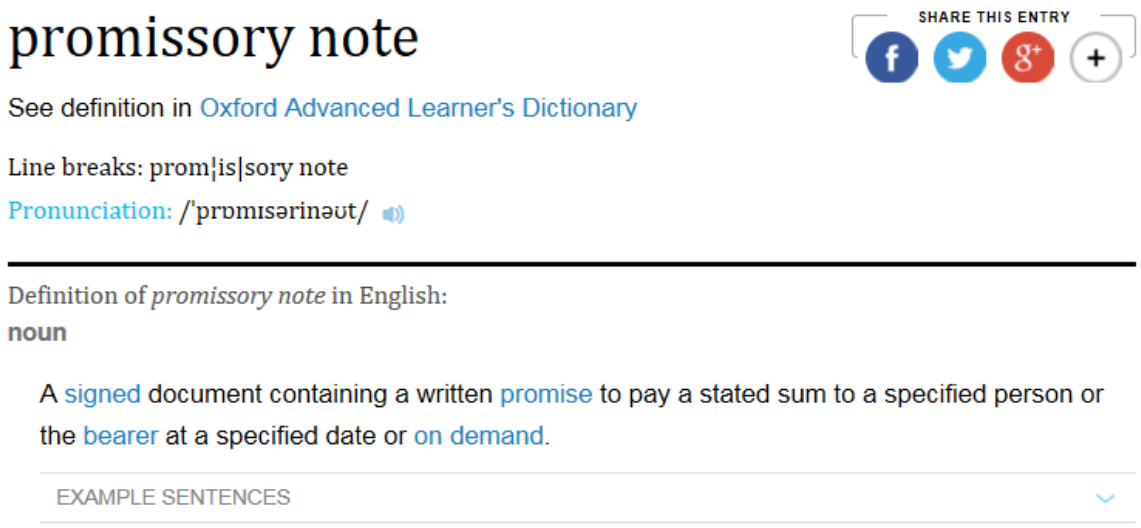


Figure 5.49: *Oxford Dictionaries* online definition for ‘promissory note’

There is no indication of what kind of document it might be and one might well need to follow the hyperlinks to find out what is meant by ‘the bearer’ or ‘on demand’. *Merriam-Webster*’s entry includes a simpler definition, but also an equally complicated one, with no explanation as to why there are two, and what the difference is, as Figure 5.50 demonstrates.

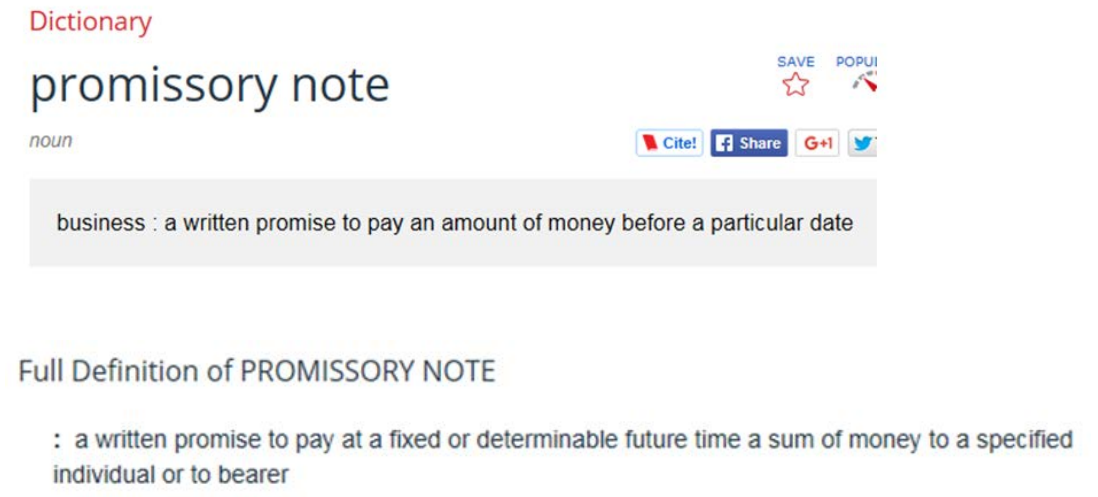


Figure 5.50: *Merriam-Webster* entry for 'promissory note'

It is unclear whether the definition marked 'business' is just a shorter version, with 'business' as a register marker, whether it is taken from a 'business' version of the dictionary, or for some other reason. In 2016, the definitions are distinguished as 'full' and 'simple', and entries appearing in other *MW* dictionaries are listed as such in the drop down menu when you initially input your search word. In 2014, however, it was a confusing dictionary to work with, as it had many multiple definitions that were not clearly distinguished. I can only imagine that the dictionary was in the process of being updated (as has happened with several of the websites used during this research project).

This coupled with the lack of information in *MW* entries in general (see 5.3.2) all suggest that *MW* is simply a less comprehensive dictionary than the others. Whether this is in any way related to its position as a 'corpus-informed' dictionary as opposed to a 'corpus-based' one is unclear, since much less information on the workings of the *MW* fleet of dictionaries is available than that on the Oxford suite. Full understanding of the *Merriam-Webster* dictionary, including its approach to neologisms, would require a dedicated piece of research which falls outside the scope of this study.

Wiktionary's definition of 'promissory note' is much clearer than any of the others and this appears always to have been the case. The definition was the same in July 2014 as

when it first entered in January 2007, and although there have been many changes recorded in the Revision History, these appear to have been to the infrastructure of the page, (including the addition of extra components such as register/style/attitude labels and synonyms) and to the number of translations. In fact 'promissory note' is translated into 18 different languages, including Japanese, Polish, Latvian, Russian and Serbo-Croatian. This perhaps gives some indication of the international understanding of and need for this word in this economic climate, although in that case I would have expected more appearances in the media than the 32 identified in the *NTON* database. I would also have expected higher annual usage, since in *NTON* use of 'promissory note' never rises above seven per year (in 2011). It may however simply be that 'promissory note' has been targeted more than other words by the bots which are automatically roving through *Wiktionary* adding translations.

Wiktionary's definition of 'promissory note' is shown in Figure 5.51.

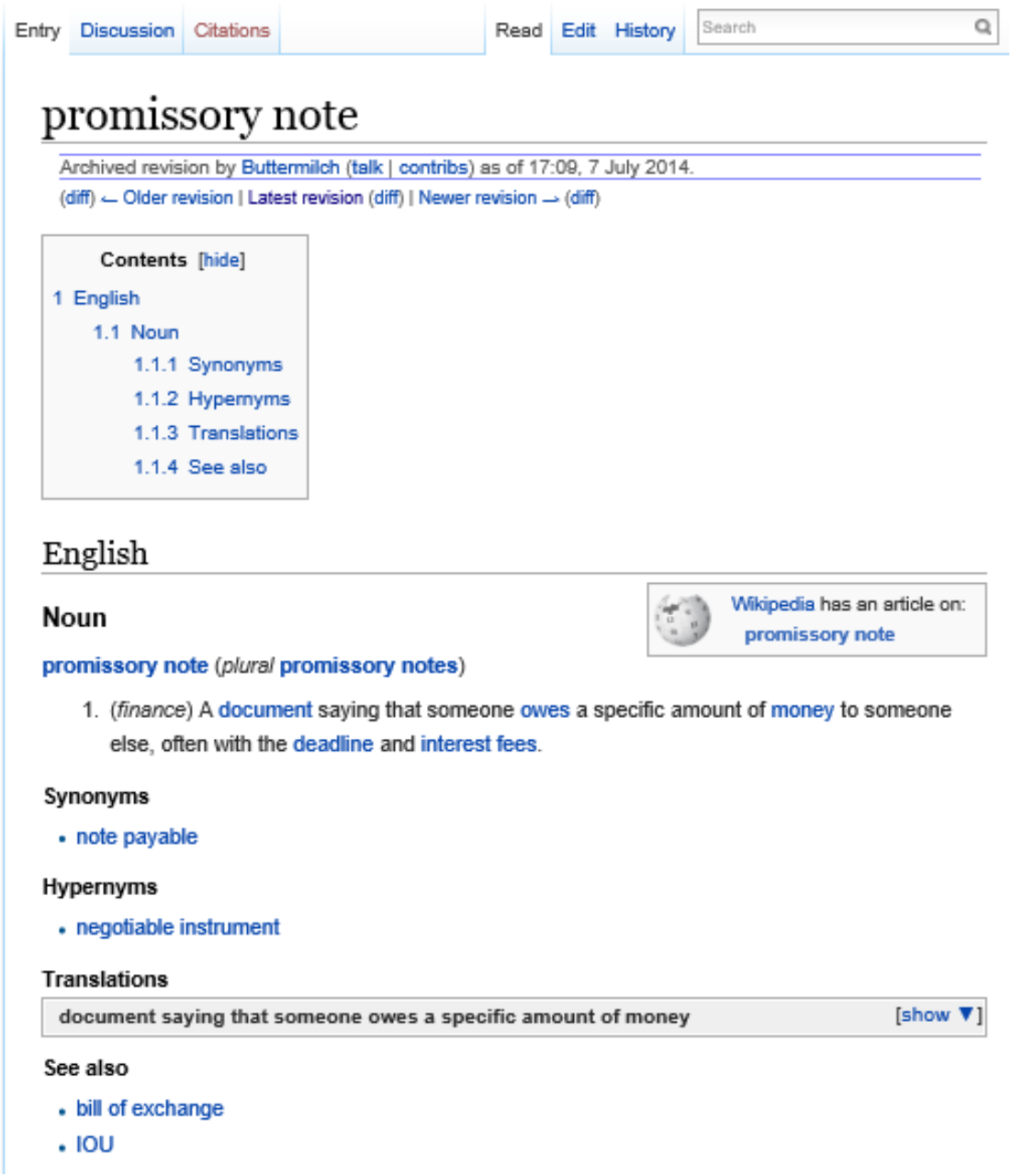


Figure 5.51: *Wiktionary* entry for 'promissory note'

Wiktionary's definition, then, makes it clear that the document concerns one person owing money to another, that it must be repaid by a pre-specified date and that interest will also be due. This entry provides much more information than do any of the more complex definitions, not just in the number of additional components, but in the language itself. Use of the word 'deadline' gives the *Wiktionary* entry a feeling of certainty that is absent from the other, vaguer, definitions, which use 'on a specified date' and 'on demand'. Indeed the expert-produced dictionaries use wording very similar to that found on British bank notes: on the side of the note featuring the image

of the Queen, under 'Bank of England', each note carries the statement 'I promise to pay the bearer on demand the sum of' and then the denomination of the note (£5, £10 and so on) appears beneath.

It is worth noting that the *NeoCrawler* definition is even more complex than that used by the expert-produced dictionaries, and reads very much as if it has been lifted directly from a financial document: 'A negotiable instrument, wherein one party (the maker or issuer) makes an unconditional promise in writing to pay a determinate sum of money to the other (the payee), either at a fixed or determinable future time or on demand of the payee, under spec' (*NeoCrawler* list, Ludwig-Maximilians Universität n.d.).

It is never made entirely clear in Kerreman's thesis how definitions for the neologisms generated and tracked by the *NeoCrawler* were arrived upon. However it appears they were taken from:

a) comments found about the new words during an early, pilot stage of the development of the program (soon rejected) which involved 'standard web crawling with metalinguistic markers as search strings'(Kerremans 2012: 63-4)

b) definitions made up as part of a survey investigating the conventionalization of English neologisms (Ibid: 158)

c) the *Urban Dictionary*[140] (Ibid).

This would explain why some of the *NeoCrawler* definitions are so poorly written. If we compare them with the *Wiktionary* definitions from the other dictionaries we can see how much more information the latter provide (see Table 5.14).

[140] http://www.urbandictionary.com/

Neologism	*NeoCrawler* Definition	Dictionary Definition
bankster	a person in the financial service industry who grows rich despite the continued impoverishment of those who depend on their services, and despite their apparent inability to succeed in business without constant government assistance	a criminal banker, often used in the plural as an aspersion against bankers in general (W*iktionary*)
cyber-bullying	the use of internet and mobile phones to send embarrassing or hurting [sic] messages	the use of electronic communication to bully a person, typically by sending messages of an intimidating or threatening nature (O*xford Dictionaries* online)
sovereign debt	a debt instrument guaranteed by a government; a bond	the amount of money outstanding that was borrowed by a government in order to finance expenditure not covered by taxation (*Wiktionary*)

Table 5.14 *NeoCrawler* definitions compared with definitions in the other dictionaries studied here). (*NeoCrawler* list, Ludwig-Maximilians Universität n.d., *Wiktionary* 2014 and *Oxford Dictionaries* online 2014)

The *NeoCrawler* definition for 'bankster' is extremely (unnecessarily) complicated. A full comparison of *Wiktionary* and expert-produced definitions of this term is conducted later in this section, therefore I will not go into further detail here.

The definition for 'cyber-bullying' is grammatically incorrect ('hurting' should read 'hurtful' or 'harmful') and this time the definition is lacking in detail, as is the definition for 'sovereign debt', which is also discussed in further detail later in this section.

One thing that should be noted here, however, is the issue of spelling variants. It was explained in 4.3.2.1 that neologisms which could be spelt either as a single word, a hyphenate or a two-word term were searched individually during the Media Tracking process, and all of these results were compounded whilst analysing the results. In terms of dictionary definitions, all bar one of these terms was spelt as a single-word-term in all of the dictionaries under study. 'Cyberbullying' however, was hyphenated in the *OED*. None of the dictionaries offer the hyphenated form as an alternative, and *OED* does not offer the unhyphenated form. I can find no reason for this anomaly in the *OED*, and can in fact only imagine that it was an error on the part of the lexicographer. Perhaps in a later update it will be corrected.

'BOGOF' is another term which can be written in multiple ways: 'bogof', 'BOGOF' and 'Bogof' (although these variants made no difference during Media Tracking). Several of

the dictionaries provide both upper and lower case versions of 'BOGOF', including *Wiktionary*[141].

The *NeoCrawler*'s definition for 'BOGOF' is notable for being better than those provided by most of the expert-produced dictionaries. *ODE* and *ODO* both simply define BOGOF as 'buy one get one free' (a synonymous definition), the term for which the initials stand. There is no explanation for what that actually means. The *NeoCrawler* explains that 'BOGOF' is an 'advertising strategy that entices people to buy a product and get one for free' (*NeoCrawler* list, Ludwig-Maximilians Universität n.d.). *Wiktionary* and *OED* both also explain that 'BOGOF' is a promotional device.

In the interests of continuity, having compared dictionary definitions for 'promissory note', it would be interesting now to consider how these compare with newspaper uses of the term. When we do so, we find that a term which appears as one of the cross-reference terms in the *Wiktionary* dictionary entry (but is mentioned in no other dictionary entry) – 'IOU' – appears as a collocate of 'promissory note' in numerous newspaper articles in my database, *Neologism Tracking in Online Newspapers* (*NTON*).

When we extract Collocation Candidates for 'promissory note' from Sketch Engine, 'IOU' is one of the top six collocates (excluding grammatical words such as 'a', 'the' and punctuation markers). It is interesting also that 'promissory note' stands out from most of the other comparisons made between dictionary definitions and media uses of neologisms in that in the samples found for 'promissory note', it is in most cases clear what the term means. For example in Figure 5.52, a reader could look at the concordance lines, and the *Wiktionary* definition, and immediately connect the two.

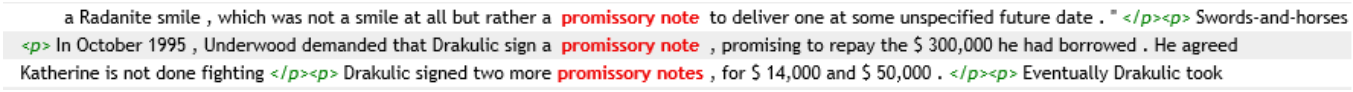


Figure 5.52: Sample uses of 'promissory note' from *NTON* database, *Independent*, 2007; *Mail,* 2014

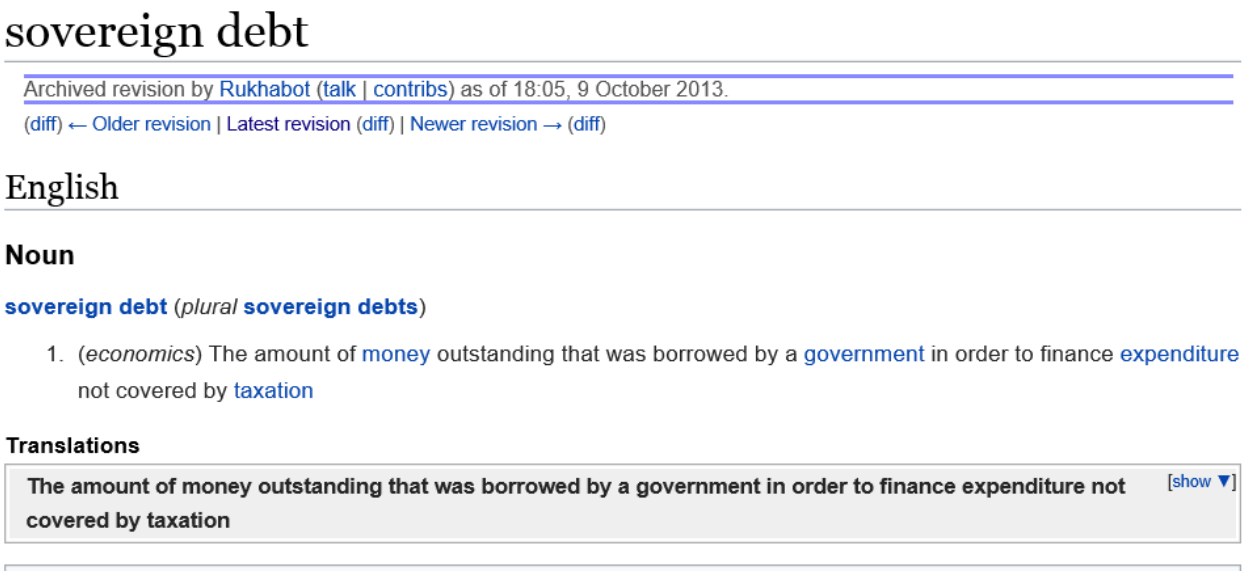
It would be slightly less straightforward to connect the other dictionaries to these examples because of the more complex language used in their definitions. None of the other dictionary entries carry any kind of component which would aid this process, not

[141] https://en.wiktionary.org/wiki/bogof

even a ‘domain label’ indicating that the word is generally used in the context of finance, money, business or banking (*Wiktionary* uses the ‘finance’ label).

If we briefly consider Research Question 1, we can see that we have learnt from this study that *Wiktionary*’s definitions are of a higher quality than those of the other dictionaries, in terms of the clarity and amount of information provided. This has been shown by examination of, for example, ‘promissory note’, ‘bankster’ and ‘BOGOF’. This also further confirms *Wiktionary*’s responsiveness to neologisms, in that it most closely reflects the real-world usage of the word as evidenced in the media.

Sovereign debt’ is a DDEB2 word from a similar field as ‘promissory note’ (finance). It appears only in *ODO* and *Wiktionary*, but it appears 1,244 times in the newspapers used for Media Tracking (see 5.4) and 925 in the *Oxford English Corpus* (*OEC*) (based on research findings derived from the *Oxford English Corpus*, Oxford University Press). As with ‘promissory note’, the expert-produced dictionary definition here (*ODO*) is complex and requires an understanding of financial issues. The *Wiktionary* definition effectively explains these issues, as shown in Figure 5.53.


sovereign debt

Archived revision by Rukhabot (talk | contribs) as of 18:05, 9 October 2013.

(diff) ← Older revision | Latest revision (diff) | Newer revision → (diff)

English

Noun

sovereign debt (*plural* **sovereign debts**)

1. (*economics*) The amount of money outstanding that was borrowed by a government in order to finance expenditure not covered by taxation

Translations

The amount of money outstanding that was borrowed by a government in order to finance expenditure not covered by taxation [show ▼]

Figure 5.53: *Wiktionary* entry for ‘sovereign debt’

As discussed earlier in this section, the *NeoCrawler* definition is even shorter than the *Wiktionary* one, and as a result, fails in its purpose to explain the meaning of the term: 'a debt instrument guaranteed by a government. A bond spec' (*NeoCrawler* list, Ludwig-Maximilians Universität n.d.). Thus of all the definitions in this study, the *Wiktionary* definition is the most detailed.

As with 'promissory note' (Figure 5.51 above), the *Wiktionary* definition for 'sovereign debt' is more detailed and comprehensive than that in *ODO,* with simpler language and more familiar terms. While the entry for 'sovereign debt' as a whole is not as comprehensive as the one for 'promissory note', it does provide a domain label showing that the word is generally used in an economics context; *ODO* does not.

See Table 5.15 for a summary showing where *Wiktionary*'s definitions are more detailed or its entries more comprehensive than those of expert-produced dictionaries. This table records the elements in *Wiktionary* entries which set it apart from its competitors.

Neologism	*Wiktionary* Definition	'Stand-out' Element	Additional Notes
acedia	1. spiritual or mental sloth 2. apathy; a lack of care or interest; indifference 3. boredom	actual definition provided, not just cross-reference to entry for 'accidie'	
bankster	(informal, derogatory) a banker who is seen as criminally irresponsible, or as extorting bailout money from the taxpayers	clarity of definition through idea of 'criminality'	
bogof	1. (chiefly Britain) buy one, get one free (a retail promotion in which consumers may purchase two items for the usual price of one 2. at item promoted in this way	definition makes clear how this retail promotion works	
conurbation	a continuous aggregation of built-up urban communities created as a result of urban sprawl	definition: urban sprawl	
cyberbullying*	n/a		noun added on 19 May 2015
cyberchondriac	a hypochondriac who researches his/her potential medical condition on the internet	clarity of definition through use of 'hypochondriac'	
diabesity	(pathology) obesity and diabetes in the same patient, especially when the obesity had a causal influence of [sic] the diabetes	domain label: pathology clarity of definition showing causal influence	
earworm	a tune that is stuck in one's head, especially an unwanted or repetitive one	clarity of definition through use of 'unwanted'	
e-tailer	a company that does business via electronic media, especially via the internet		should mention that it relates to selling, or goods/products
e-waste	discarded electric and electronic equipment		
floordrobe	(humorous) clothing strewn on the floor	register/style/attitude label: humorous	no other definition to compare to
frenemy	1. (humorous) someone who pretends to be your friend, but is really your enemy 2. (humorous) a fair-weather friend who is also a rival	register/style/attitude label: humorous clarity of definition through use of 'pretends' and 'fair-weather'	

cont/d…

gendercide	the killing of people because of their gender		no other definition to compare to
globesity	the worldwide obesity epidemic		no other definition to compare to
greenwashing*	n/a		*Wiktionary* does contain noun 'greenwash'
hubristic	1. of or relating to hubris; overly arrogant 2. displaying hubris (as a personality characteristic)	clarity of definition through use of 'personality characteristic'	*OED* definition different to all others: 'insolent, contemptuous'
hyperlocal	(chiefly journalism and blogging) related to a very small area, smaller than normally considered local	domain label: chiefly journalism and blogging	
promissory note	(finance) a document saying that someone owes a specific amount of money to someone else, often with the deadline and interest fees	domain label: finance clarity of definition through 'owes'; 'deadline'; 'interest fees'	
rewilding	n/a		only entry appears in *OED*
sovereign debt	(economics) the amount of money outstanding that was borrowed by a government in order to finance expenditure not covered by taxation	domain label: economics clarity of definition: 'expenditure'; 'taxpayers'	
superphone	(informal) a remarkably advanced telephone		no other definition to compare to
tenebrous	dark and gloomy		
upskill	1 (transitive) to teach someone additional skills, especially as an alternative to redundancy 2 (intransitive) to acquire such additional skills	extra information in definition: 'alternative to redundancy'	
warrantless	(of a search, arrest or the like) performed without a warrant	clarity of definition: 'without a warrant'	
waterboarding	a type of torture technique in which the victim is immobilized, has towels or rags wrapped over their face, and has water poured onto them, which simulates the sensation of drowning		should mention that the detainee is strapped down
wellderly	old people who are in good health		no other definition to compare to

*includes spelling variants

Table 5.15: Elements of *Wiktionary* definitions and/or entries as at 31 August 2014

The items in the 'Stand-Out Elements' column of Table 5.15 show the areas in which the *Wiktionary* entries stand apart from all four of the expert-produced dictionaries. The *Wiktionary* entries are the most detailed and/or the most comprehensive for all but seven of the 26 neologisms appearing in one or more dictionary.

The seven less detailed *Wiktionary* entries are:

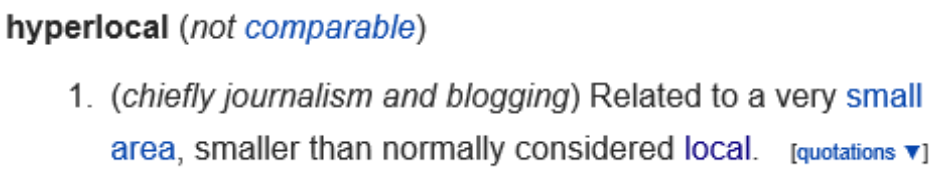
- 'cyberbullying' (DDEB2) – not yet in *Wiktionary* as a noun
- 'Greenwashing' (DDEB3) – not yet in *Wiktionary*, although 'greenwash' is present
- 'e-tailer' (DDEB3) – should mention 'selling' and 'products/goods'
- e-waste (DDEB3) – needs examples
- 'tenebrous (DDEB3) – should include examples
- 'rewilding' (DDEB3) – not yet present in *Wiktionary*
- 'waterboarding' (DDEB3) – *Wiktionary* entry should mention that the detainee is strapped down

Of the remaining 19, often it is the presence of a single word or phrase which gives the *Wiktionary* definition the edge over its 'corpus-based' and 'corpus-informed' competitors. For example although the definitions for 'cyberchondriac' used in the Oxford suite of dictionaries are perfectly understandable, the addition of the word 'hypochondriac' in the *Wiktionary* version gives the definition, and the word it is seeking to describe, an air of familiarity. It positions 'cyberchondria' (a new condition) within the well-recognised 'hypochondria'. 'Hypochondria' does not appear in the same immediate context as 'cyberchondriac' in either the *NTON* database or the *OEC* (based on research findings derived from the *Oxford English Corpus*, Oxford University Press), and it is included in neither Collocation Candidate lists nor Word Sketches for 'cyberchondriac', yet the similarities to the morphology of 'hypochondriac', and the

familiarity of that term make 'cyberchondriac' more understandable in *Wiktionary* than in the other dictionaries which do not mention 'hypochondriac'.

The same is true, I believe, of the use of the concept of 'criminality' in *Wiktionary*'s entry for 'bankster'. While the *ODO* definition talks about 'banksters' being viewed as 'profiteering' and 'dishonest', *Wiktionary* goes so far as to state that they are seen as 'criminally irresponsible' and 'extorting' money. This is a much harsher viewpoint, and reflects the formation of the word, as a blend of 'banker' and 'gangster' (see 5.4.1.1 regarding word formation processes).

Finally we turn to 'hyperlocal'. All of the dictionaries make clear that the term indicates an area much smaller than would normally be designated 'local'. *Wiktionary*, however, is again more precise, adding a subject label stating that the term applies mainly to 'journalism and blogging', see Figure 5.54.


hyperlocal (*not comparable*)

1. (*chiefly journalism and blogging*) Related to a very small area, smaller than normally considered local. [quotations ▼]

Figure 5.54 *Wiktionary* entry for the neologism 'hyperlocal'[142]

To summarise this discussion of the comparison between definitions from different dictionaries, in response to Research Question 1, we can conclude that as with the other dictionary components discussed in 5.3.2, *Wiktionary* appears more comprehensive than expert-produced dictionaries. It provides clearer and more accessible dictionary definitions than those found in either 'corpus-based' or 'corpus-informed' expert-produced dictionaries, and this seems to be as a result of *Wiktionary*'s collaborative nature. Contributors with knowledge in a wide range of fields are able to make entries both more detailed and less complex, often through the addition of a single word, such as 'hypochondriac' in the entry for 'cyberchondriac'.

[142] https://en.wiktionary.org/w/index.php?title=hyperlocal&oldid=26825811

5.3.5 Conclusion: Lexicographical Perspectives

Through the course of this discussion, in response to Research Question 1, *Wiktionary* has been shown to be the most comprehensive of the dictionaries under study here, in terms of its responsiveness to neologisms, the number and quality of dictionary components its entries contain, and the quality of its definitions when compared to those in expert-produced dictionaries.

Wiktionary achieves the higher level of detail in its new word entries in part because of the flexibility it experiences through not being bound by the conventions of standardised dictionary structures. Hence it can include not only standardised dictionary components such as examples, pronunciation guidance and domain labels, but also elements such as inclusion dates and discussion forums. Detailed, attestational illustrative examples can be drawn from sources that would not be permitted in traditional dictionaries, such as television programmes and music covers, by *Wiktionary* contributors who comprise the other key factor in *Wiktionary*'s success. They build upon each other's work, adding to and expanding dictionary entries so that much more information is ultimately included than could have been supplied by one or two individual operators. These contributors are also responsible for definitions appearing in *Wiktionary* new word entries, definitions which have been shown to be of higher quality than their 'traditional' competitors. These contributors (apparently unwittingly) use the same defining styles as do expert lexicographers, and their definitions contain just enough extra information, just enough less information or just a simple clarifying word or two to make the definition appear more accessible and understandable for users. All of this work by these contributors, coupled with the immediate updating of the site, means that *Wiktionary* not only provides more detail in its neologism entries, and higher quality definitions, but is also significantly more responsive to neologisms than are expert-produced dictionaries. As was shown with the case of the 'frenemy' in 5.3.3, during the course of just 24 hours, an entry can go through an entire 'proof of qualification' period, with some contributors deeming it unsuitable as an entry and putting it up for deletion, while others respond by finding and presenting the necessary attestational evidence needed to prove that it meets *Wiktionary*'s inclusion criteria and does therefore merit entry in the dictionary.

The fact that *Wiktionary* utilises the same types of examples, as well as defining styles as entries in both 'corpus-based' and 'corpus-informed' expert-produced dictionaries means that it is competing on an even more 'like-with-like' basis than was expected at the start of the project.

5.4 Media Tracking: Neologism Use and Behaviour in UK National Newspapers

In this section, I present and discuss findings from the Media Tracking element of this study, which was used to trace the use and behaviour of the 34 neologisms, whilst at the same time comparing the new manual data collection methodology devised during this project with the most recently written up automated system of a similar nature, the *NeoCrawler*. The objectives of this part of the study were thus two-fold:

1. To track neologism appearances in UK news media in order to compare usage and behaviour in different newspapers at different stages in the neologic life-cycle
2. To consider whether neologism use and behaviour in the media can be best explored through the use of new manual or existing automated corpus data collection techniques.

In the course of this, I draw conclusions in answer to Research Questions 2 and 3:

> **Research Question 2 –** *What can be discovered about the 'neologic life-cycle' of selected neologisms in UK national newspapers between 2000 and 2014?*
>
> **Research Question 3 –** *In the context of data collection for context-rich, genre-specific web-based corpora, is the proposed new manual methodology more or less appropriate and effective in tracking neologism use and behaviour than the automated methods of the kind used by the* NeoCrawler*?*

Before beginning, I first offer a brief summary of my findings, which outlines what can be discovered about the neologic life-cycle of neologisms in UK national newspapers, and which methodological approach is most appropriate and effective in tracking neologism use and behaviour.

The Media Tracking process used to achieve Objective 1 above provided a range of useful information on the newspapers most likely to include articles containing neologisms. For example it showed that *The Guardian* includes more neologism appearances across the entire neologic life-cycle (the 14-year period of this study) than any other newspaper. However while this was the case, there was a distinction between the newspaper most likely to include better established neologisms (stage DDEB3 of the neologic life-cycle) (*The Guardian)* and the one most likely to include emerging neologisms (those with the beginnings of a 'dictionary track record': stage DDEB2 of the neologic life-cycle) (the *Independent)*. The study also showed that words entering a new phase in their life-cycle can present as 'new' when, for example, they enter *Wiktionary* for the first time (for example members of stage DDEB3 which are categorised here as 'reincarnated').

In comparing the manual methodology devised during the course of this project with the *NeoCrawler* automated system, the study showed that the former is more appropriate for this kind of genre-specific study, since it allows for the collection of key contextual information. In this case, the publication date of articles was crucial in order to present neologism newspaper appearances in the correct order. Excluding unwanted information was also key however, and again the manual methods used here meant that articles which could potentially have skewed results, for example those paid for through external sponsorship and not written by professional journalists, were not included.

5.4.1. Neologisms and Word Formation Processes

As mentioned in 1.2.2.1, it is important in a study of neologisms to keep in mind the word formation processes which have created them. After presenting the morphology of several of the words selected for inclusion in this study in 5.3.2, in the following subsections I present findings on the distribution of these different categories of morphology against appearances in UK national newspapers. These findings allow me to draw conclusions in answer to Research Question 2:

What can be discovered about the 'neologic life-cycle' of selected neologisms in UK national newspapers between 2000 and 2014?

5.4.1.1 Derivational Word Formation Processes in the NTON *Database*

All of the neologisms in Dictionary Date of Entry Batches 1 and 2 (DDEB1+2) of the *NTON* database were created through derivation, while 66.8% of DDEB3 words were also derived from existing terms, as we can see in Table 5.16. This shows the relative frequency of word formation processes across the two datasets (DDEB1 and 2 are generally combined together as in this case, to form a single overarching dataset based upon 'newness' of neologisms, centring around the cut-off date on the neologic life-cycle for least well-established words of September 2008).

Dictionary Date of Entry Batch	**Word Formation Processes**	**Percentage of DDEB Total Neologisms**
DDEB1+2	Affix	12.6%
	Blend	41.1%
	Compound	46%
DDEB3 `	Acronym	4.88%
	Affix	4.46%
	Blend	11.93%
	Calque	4.95%
	Compound	45.45%
	Reincarnated	28.33%

Table 5.16: Percentages of Word Formation Processes in *NTON* database

The derivational words here are comprised of acronyms ('blends made up of initial letters'), affixes (appearing at the beginning and ends of words), blends (parts of words joined together) and compounds (entire words joined together) (Carstairs-McCarthy 2002: 21, 59, 65). The remaining third are loan words ('calques', for example 'earworm', which according to *Wiktionary* is borrowed into English from German[143]) and words

[143] https://en.wiktionary.org/wiki/earworm

which do not fall neatly into any currently recognised word formation method. These are termed here as 'reincarnations' (see 5.4.1.2). Table 5.17 shows the breakdown of neologisms in *NTON* by word formation type.

Neologism	WFP	Components	Total Neologism Appearances
acedia	reincarnated		4
bankster	blend	bank+gangster	3
bogof	acronym	buy one get one free	141
buzz marketing	compound	buzz+marketing	13
cold peace	compound	cold+peace	22
conurbation	reincarnated		327
cyberbullying*	blend	cyber+bullying	1196
cyberchondriac	blend	cyber+chondriac	12
diabesity	blend	diabetic+obesity	11
earworm*	calque from German		143
e-tailer	blend	electronic+retailer	112
e-waste	blend	electronic+waste	49
floordrobe	blend	floor+wardrobe	6
frenemy	blend	friend+enemy	47
gendercide	blend	gender+genocide	19
globesity	blend	global+obesity	10
greenwashing*	blend	green+whitewashing	118
hubristic	reincarnated		441
hyperlocal*	affixation	hyper+local	292
newer markets	compound	newer+markets	22
open education	compound	open+education	8
predatory lending	compound	predatory+lending	31
promissory note	compound	promissory+note	33
rewilding**	affixation	re+wilding	92
round pound	compound	round+pound	9
sodcasting	blend	sod+casting	4
sovereign debt	compound	sovereign+debt	1244
superphone*	compound	super+phone	31
tablet computing	compound	tablet+computing	24
tenebrous	reincarnated		47
upskill	affixation	up+skill	41
warrantless	affixation	warrant+less	88
waterboarding*	compound	water+boarding	1281
wellderly	blend	well+elderly	19

* includes spelling variants: hyphenated and two-word version

** includes spelling variant: hyphenated

Table 5.17: Neologism Word Formation Processes

Neologisms appearing in DDEB1+2 were created either through affixation, blending, or compounding. Of these, 46% were created through compounding; in DDEB3 the figure is 45.45%. As discussed in 1.2.2.1, within the study of morphology these word formation processes, as well as the acronyms and loan words found in DDEB3, are considered standard methods of creating new words (Minkova and Stockwell 2009: 11-19).

When we bring all of the neologisms from the two Dictionary Date of Entry Batches (DDEB stages of the neologic life-cycle) together, looking at actual instances of usage, we see that, as with the individual datasets, the greatest number were created through compounding (2,718 uses in the database as a whole; 1,404 in DDEB1+2 and 1,314 in DDEB3). This is followed by 1,606 blends (1,261 plus 345), as shown in Figure 5.55 which brings together all neologisms by word formation type.

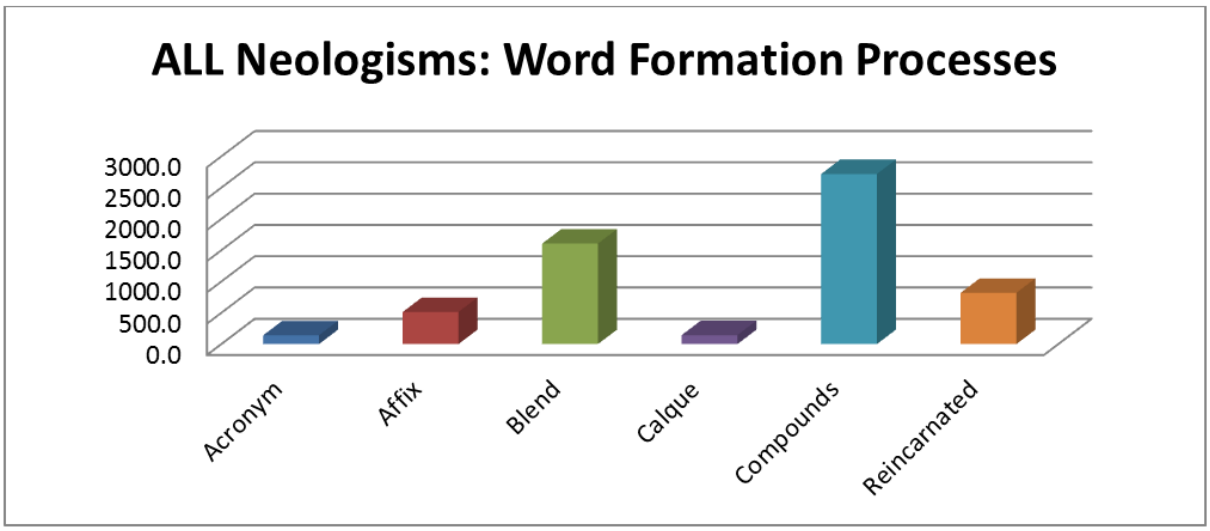


Figure 5.55: Spread of Word Formation Processes across all neologisms

Compounding is widely considered to be the most popular method of creating new words (see for example Minkova and Stockwell 2009: 9-11). However when we examine blending (the next most common form of derivation in both datasets) we see that in the *NTON* database, for the neologisms most recently included in a dictionary, or yet to enter a dictionary at all (DDEB1+2) the difference between compounds and blends is significantly less than for DDEB3, neologisms which had already entered a dictionary by

August 2008 (41.6% of DDEB1+2 neologisms are blends, as compared with 11.93% of DDEB3 terms). Figures 5.56 and 5.57 provide a graphical representation of Table 5.17, breaking down datasets DDEB1+2 and 3 by percentage of word formation process.

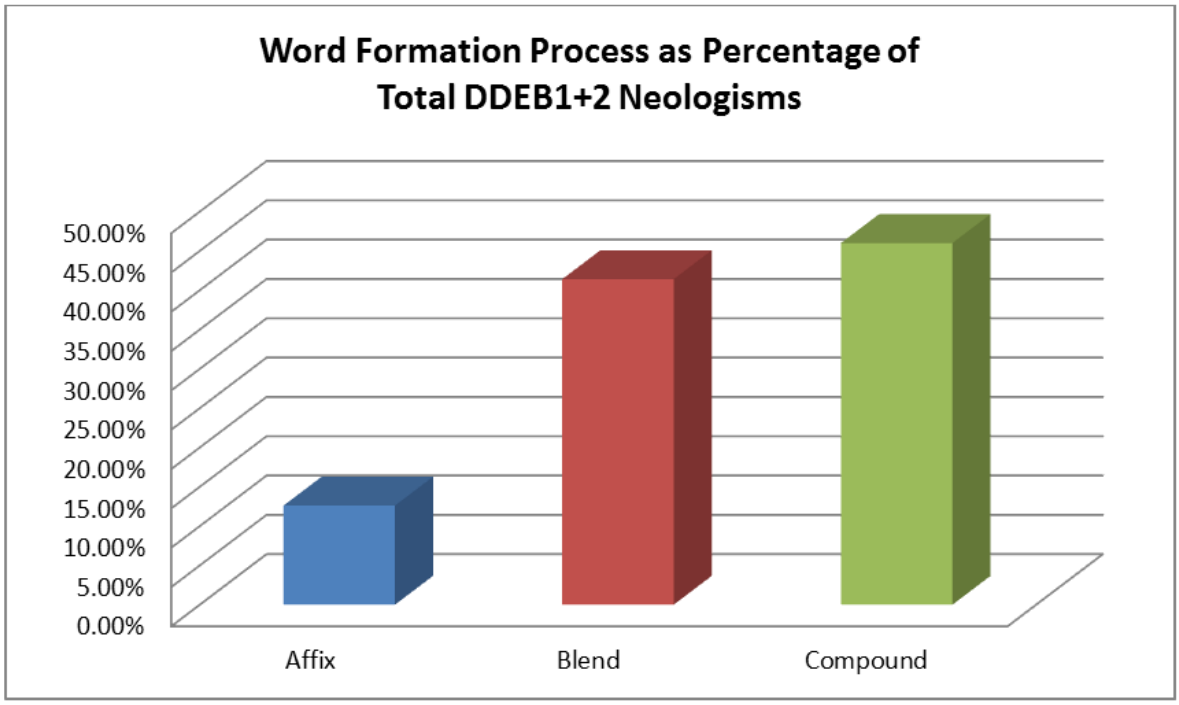


Figure 5.56: Percentage of Word Formation Processes across DDEB1+2 neologisms

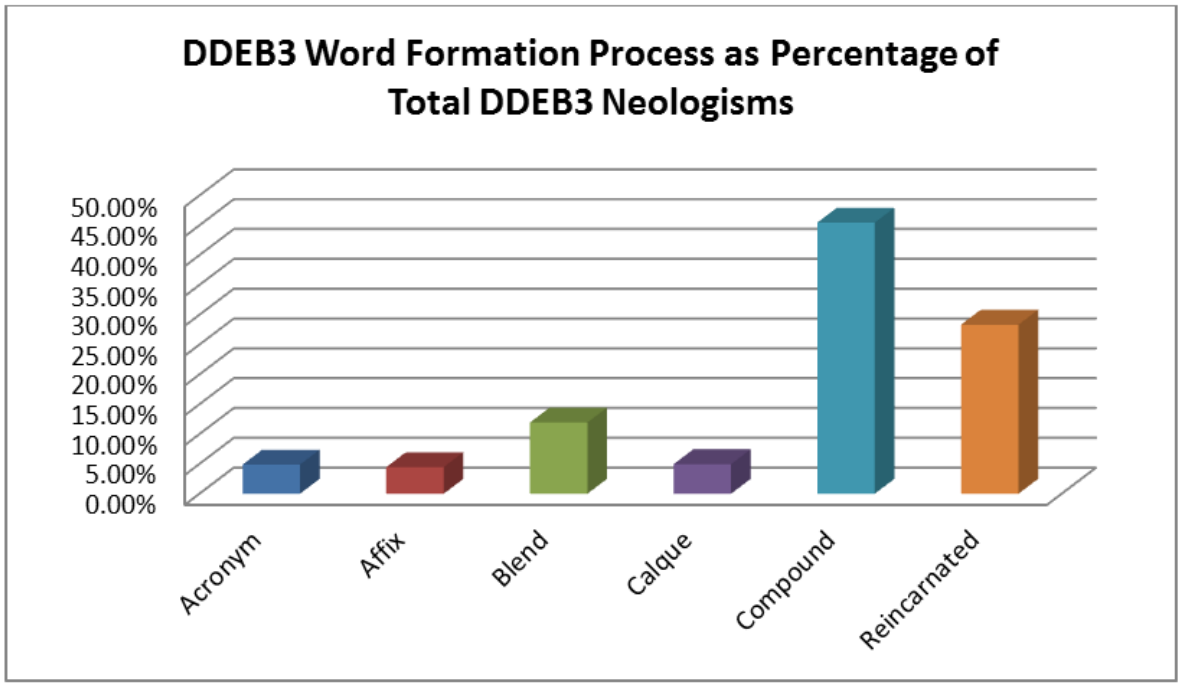


Figure 5.57: Percentage of Word Formation Processes across DDEB3 neologisms

This evidence may support Lehrer's argument in her 2003 paper 'Understanding Trendy Neologisms' that blending is gaining popularity as a method of new word formation. The newer words in this study are more likely to be blends, although as Lehrer points out blending is not in itself a new process. However blends can be more difficult to understand on first viewing because the reader/listener must interpret the sections of words that have been brought together and then understand the synergistic new terms they create (Ibid: 369, 371-2).

Returning to Research Question 2, then, we can say that from examining the selected neologisms in UK national newspapers between 2000 and 2014, across the neologic life-cycle, compounding was the most popular method of word formation. This is believed to be the case in word formation in general, and hence the current study appears to reflect wider trends (see for example Minkova and Stockwell 2009: 9-11). However in the most recent stages of this study's neologic life-cycle, only two other word formation processes were used, whereas amongst earlier, better established DDEB3 terms, an additional five were employed, including the non-standard 'reincarnated' process (see below). It is not clear why this differential appears, although it seems likely that the blending of DDEB1+2 is taking the place of the more diverse processes apparent in DDEB3. It would be useful to conduct a similar study in a further 10 to 15 years to see how these new patterns are developing.

5.4.1.2 Non-Standard Word Formation Processes in the NTON *Database*

In this section I consider Research Question 2 in light of the four 'reincarnated' terms identified in 5.3.1:

> *What can be discovered about the 'neologic life-cycle' of selected neologisms in UK national newspapers between 2000 and 2014?*

Of the 5,940 neologisms in the *NTON* database, 13.8% are what I consider to be of non-standard morphology. These are 'acedia', 'conurbation', 'hubristic' and 'tenebrous' from DDEB3. Accounting for 28.33% of DDEB3 words, it is believed that these terms experienced extensive use in the past, but at some point fell out of favour, such that when they began to gain popularity once more, as found during the *NeoCrawler's* work

within Google Blogs, they were deemed to be 'new' (Kerremans 2015). I originally hypothesised that this was because these terms simply did not arise in the *Google Blogs* environment. However I have since come to believe that the context in which the words appeared is not relevant. What is relevant is the internal dictionary which was used by the *NeoCrawler* to assess whether or not a word was a potential neologism. This was made up of *Wikipedia* and Google N-Grams (Ibid: 81). It is my belief that these words were included in the list of *NeoCrawler* neologisms through an error in this automated dictionary-checking process. As *Wikipedia* formed part of this dictionary, I checked for the presence of these words at that time. Had the words been included in *Wikipedia*, one would have expected them to be automatically excluded as potential neologisms. 'Tenebrous' entered *Wikipedia* in 2007 and 'acedia' in late 2006, suggesting they would indeed have been missing from the internal dictionary (whose cut-off point was before this), and therefore selected as candidate new words. 'Hubristic' and 'conurbation' were present in *Wikipedia*, however (indeed 'conurbation' had its own *Wikipedia* page dating from July 2003[144]), indicating that the *NeoCrawler* team should have discounted them, yet did not. This is a situation then, in which it appears that automated systems have introduced errors which manual methods could have avoided, since my new methodology involves 'advance exploration' of websites (see 4.5.2) which would have identified the *Wikipedia* entries.

The newspaper's websites only search back 20-30 years. I therefore looked in *The Guardian and Observer Digital Archive*[145] for examples of slumps in usage, in an effort to support my theory that these four words rose and fell in popularity over time. Old documents (from the 1900s to the 1960s) were examined to discover the points at which they dropped out of use and then reappeared. Such slumps were indeed found, however the information in the archive was discovered to be highly unreliable.

The four words entered *Wiktionary* in 2005/6, and changes in their media usage occurred around this time, suggesting that the two events were related. For the purposes of my own research, I took the decision to retain these four words (and two others) despite the fact that they had been in use for an extended period of time. Each

[144] https://en.wikipedia.org/w/index.php?title=Conurbation&dir=prev&action=history
[145] https://www.theguardian.com/info/2012/jul/25/digital-archive-notice

had undated entries in *Merriam-Webster* and were dated as having entered the *New Oxford Dictionary of English* when it was first published in 1998. Early entries were also apparent in the *Oxford English Dictionary* (*OED*), as discussed in 5.2, and shown in Figure 5.3. 'Acedia' entered *OED* in 1933, 'conurbation' in 1972, 'hubristic' in 1899 and 'tenebrous' in 1911. In addition, 'warrantless' was dated 1921 and 'upskill' was dated 1993. ('Promissory note' was undated since it is a derivative of 'promissory' (see 3.4.4 for further discussion of dates of entry into expert-produced dictionaries)).

Despite these earlier entries in expert-produced dictionaries, these four words had first entered *Wiktionary* in 2005 and 2006, an indication of 'newness', since most of the initial entries in *Wiktionary* had come from out-of-copyright dictionary, *Webster's New International Dictionary of the English Language* (Meyer and Gurevych 2012: 262) in 2002. Yet these had not been included at that stage, suggesting 'newness' a few years later. I was therefore interested to see how these words might differ from others in neologic life-cycle stage DDEB3, in terms of their use and behaviour in newspapers and in terms of their dictionary definitions

The most interesting finding was that, despite having clearly been in existence for many years, two out of these four words ('conurbation' and 'hubristic') saw definite increases in newspaper appearances around the time of their entry into *Wiktionary* (2005). Unlike 'sovereign debt' and 'cyberbullying' (see 5.4.3), I have been unable to identify external societal or cultural factors which would have led to these increases. Neither 'acedia' nor 'tenebrous' saw similar changes. Indeed 'acedia' was used only four times during the course of the study, all after 2009, while 'tenebrous' averaged 3.35 uses per year throughout the study. The number of uses did not rise above five per year until 2009 (having entered *Wiktionary* in 2005), and only in 2012 did the number rise above 10.

Despite these increases in usage, since 'acedia', 'conurbation', 'hubristic' and 'tenebrous' had clearly not been recently created, it did not seem appropriate to apply standard word formation labels to these words. I therefore coined the term 'reincarnated' to describe them, reflecting the idea that they had 'risen again'. This idea appeared to be borne out by the subsequent rising levels of media usage of 'conurbation', 'hubristic' and 'tenebrous' – as shown in Figure 5.58.

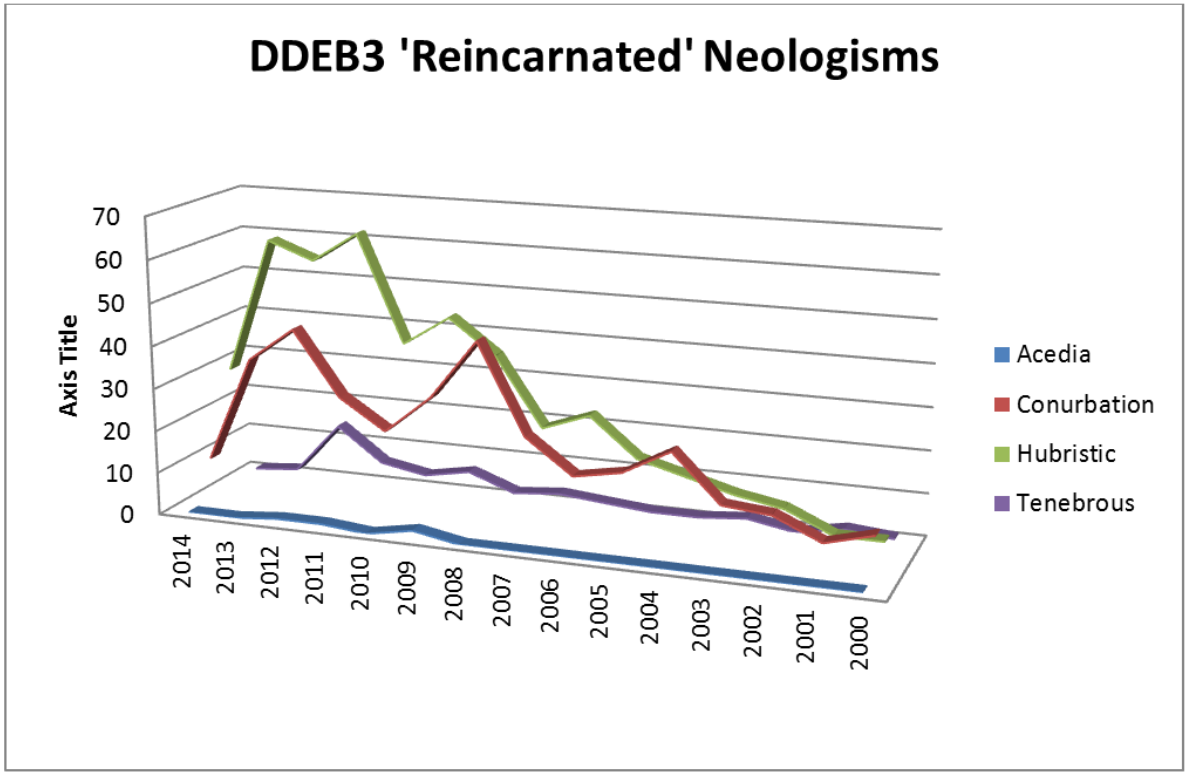


Figure 5.58: Media use of 'reincarnated' neologisms 'acedia', 'conurbation', 'hubristic' and 'tenebrous' between 2000 and 2014

Of the four 'reincarnated' words, 'hubristic' experienced the most growth, rising from four uses in 2000 to 64 in 2011. 'Conurbation' and 'tenebrous' both peaked a year later, at 43 and 14 respectively. Since then, all have gone into decline, raising the question as to whether or not these 'reincarnated' neologisms are in fact on the verge of fading from use once more, perhaps to rise again in a few more decades. A significantly longer-term research project would be required to track the true life-cycle of these words, however an ability to come and go in popular use would suggest an impressive staying power and perhaps call into question Algeo's 1993 argument that more than half of neologisms eventually disappear from use. Perhaps his 50-year study was simply not long enough to track reincarnations of former 'neologisms', and the *OED*'s practice of never removing a word once it has been accepted is in fact the right one.

The term 'acedia' did not show the same growth as the other three 'reincarnated' words, remaining unused until 2009, when it appeared just twice, both in *The Guardian*. It was used twice more during the research period, once in 2011 (*Independent*) and once in 2012 (*Guardian*); thus while it may have been experiencing increasing usage in

Google Blogs (the source for corpus texts in the *NeoCrawler*, and the reason it was included in this study), its reincarnation was perhaps not as complete as that of the other three words.

Returning to Research Question 2 specifically in light of these 'reincarnated' words then, we discover that although these four terms can be said to be significantly older than the other words in the study, having in some cases been in use for more than a century, the neologic life-cycle still applies to them, due to the way in which they come and go in the language. This is because they still meet one or more of the criteria required under the definition of 'new' adopted for this study, and they can therefore be allocated to one of the three Dictionary Date of Entry datasets (in this case DDEB3) which underpins the definition of the neologic life-cycle (see 1.1 and 3.9):

- They were (erroneously) categorised as new by the *NeoCrawler* and
- Table 5.3 in Section 5.2 shows that they all appear in all five dictionaries under study, although they only recently entered *Wiktionary* and
- Except 'acedia', none of them as yet experience consistent year-to-year usage in the four newspapers studied here.

Widening our focus once more to include assessment of word formation processes used for the creation of the remaining neologisms in this study, we see that this was undertaken in much the same way as Kerremans in her 2015 study of the *NeoCrawler* (83). That is to say, in each case, words were manually examined and assigned a word formation label. None of the 'reincarnated' words were included in Kerremans' analysis of the *NeoCrawler* neologisms, meaning it is not possible to compare her approach to these unusual words with my own. She did, however, include 'sodcasting' in her analysis, and although there is no discussion of choice of word formation label, in her Appendix 2 list of neologisms (Ibid: 246) she claims 'sodcasting' is an example of 'suffixation', with 'casting' a suffix of 'sod'. While this is morphologically sound, since 'sod' is a free morpheme and '-casting' is not (Carstairs-McCarthy 2002: 20-1), she gives no explanation of 'sod', only a definition of the word as a whole: 'the act of playing loud music on a phone or loudspeaker' (Ibid: 250). By contrast, I categorise 'sodcasting' as a

blend of ‘sod’ (‘an unpleasant or obnoxious person[146]) and ‘-casting’, from ‘broadcasting’.

This is not the only instance in which I disagree with Kerremans’ morphological analyses. While she categorises many words as blends (with which I do agree) (2015: 246-50), there are some which I would also categorise as blends that she categorises as formed by affixation. For example she lists ‘e-tivity’ (‘an online exercise or learning activity’) as ‘prefixation (with base modification)’. While I do not include ‘e-tivity’ in my study, I do include ‘e-tailer’ and ‘e-waste’, each of which I categorise as a blend: **e**lectronic+re**tailer** and **e**lectronic+**waste**. There are also inconsistencies in her categorisation: ‘cyberchondriac’ is listed as a blend, while ‘cyberdisinhibition’ is listed as prefixation. Again, I do not include the latter in my list, but I do include ‘cyberbullying’ and ‘cyberchondriac’, both of which I categorise as blends: **cyber**+**bullying** and **cyber**+hypo**chondriac**.

While I may disagree with Kerremans on some issues of morphology, both I and the *NeoCrawler* team agree that the optimal method of judging word formation labels is through manual review of words (Ibid: 83).

I go further, in saying that manual methods are preferable in general, since they enable more accurate tracking of neologism use and allow for the collection of detailed contextual information such as that discussed above on the rise and fall in use of ‘reincarnated’ words.

5.4.2 Neologisms Usage across Newspapers

In this section I present and discuss findings on the behaviour and use of neologisms in UK national newspapers, thereby continuing to explore Research Question 2, looking at the relationships revealed between particular newspapers and new words, and the differences discovered between stages on the neologic life-cycle: DDEB1+2 (neologisms not yet appearing in a dictionary, or entering between 2008 and 2014 (which for the purposes of this part of the study I combine together, since I am examining neologism use by date)), and DDEB3 (neologisms entering a dictionary between 2000 and 2008).

[146] https://en.oxforddictionaries.com/definition/sod

Table 5.18 provides a guide as to the popularity of each neologism within the media under this study. It shows the DDEBs, the number of neologism uses and the number of articles in which they appeared.

Neologism	Number neologism uses	Number articles	Dictionary Date of Entry Batch (DDEB)
acedia (n)	4	4	DDEB3
bankster (n)	3	2	DDEB2
bogof* (n)	141	99	DDEB3
buzz marketing (n)	13	9	DDEB1
cold peace (n)	22	20	DDEB1
conurbation (n)	327	299	DDEB3
cyberbullying* (n)	1196	645	DDEB2
cyberchondriac (n)	12	7	DDEB2
diabesity (n)	11	5	DDEB2
earworm (n)	143	77	DDEB3
e-tailer (n)	112	93	DDEB3
e-waste (n)	49	38	DDEB3
floordrobe (n)	6	6	DDEB2
frenemy (n)	47	38	DDEB3
gendercide (n)	19	13	DDEB2
globesity (n)	10	8	DDEB2
greenwashing* (n)	118	94	DDEB3
hubristic (adj)	441	420	DDEB3
hyperlocal* (adj)	292	185	DDEB2
newer markets (n)	22	22	DDEB1
open education (n)	8	6	DDEB1
predatory lending (n)	31	25	DDEB1
promissory note (n)	33	28	DDEB3
rewilding* (n)	92	50	DDEB2
round pound (n)	9	9	DDEB1
sodcasting (v)	4	3	DDEB1
sovereign debt (n)	1244	889	DDEB2
superphone* (n)	31	20	DDEB2
tablet computing (n)	24	23	DDEB1
tenebrous (adj)	47	47	DDEB3
upskill (v)	41	35	DDEB3
warrantless (adj)	88	70	DDEB3
waterboarding* (n)	1281	575	DDEB3
wellderly (n)	19	9	DDEB3

*includes spelling variants
DDEB1; DDEB2; DDEB3
Table 5.18: All neologisms in the *NTON* database

As noted in 5.2, the *NTON* database, built using the new methodology devised during the course of this project, comprised 4.2 million words, including 5,940 occurrences of the neologisms under investigation. *The Guardian* was responsible for 2,204 or 37.10% of these, with the next closest frequency to this being the *Mail*, with 30.67% (1,822 instances), followed by the *Independent* (25.01%, 1,486) and finally the *Express* at just 428 neologism uses or 7.20% of the total. This breakdown is shown in Figure 5.59.

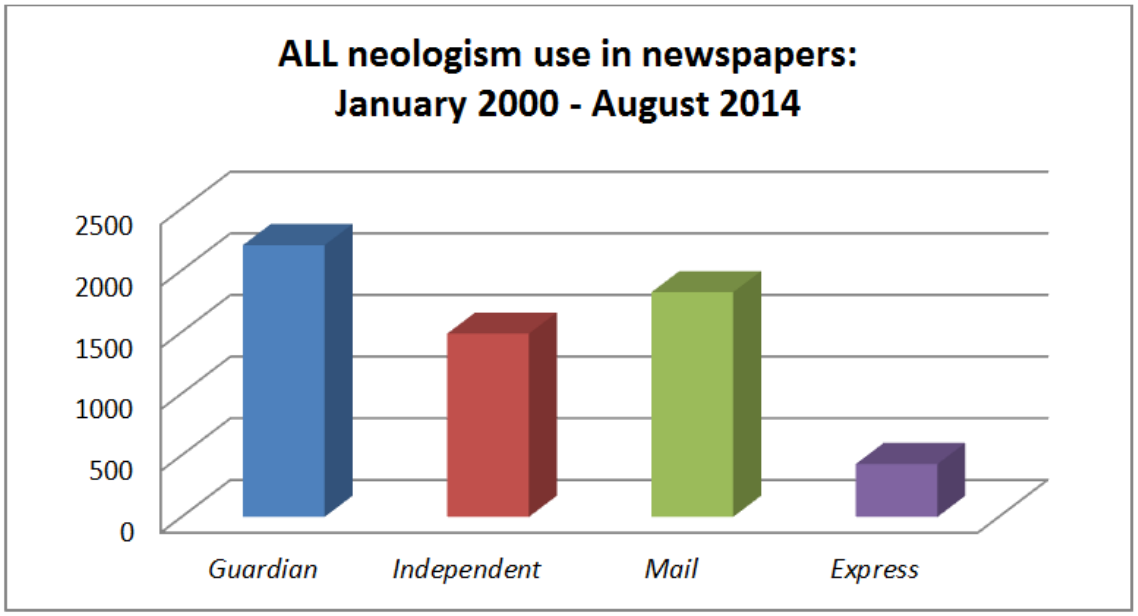


Figure 5.59: Spread of neologism usage in newspapers across the *NTON* database, January 2000-August 2014

This dominance by *The Guardian* stemmed largely from DDEB3, as it was responsible for 39.67% of DDEB3 neologism usage, as opposed to 28.91% in the *Independent* and 25.73% in the *Mail*. In DDEB1+2, *The Guardian* was pipped at the post by the *Mail*, with 34.67% and 35.36% respectively. The *Independent* carried 21.32% of neologism uses. *The Express* was the only newspaper to consistently appear at the same point in the rankings, always with significantly fewer neologism uses than all the others: 8.66% in

DDEB1+2 and 5.67% in DDEB3. The raw data for these figures is shown in Tables 5.19 and 5.20, and demonstrated graphically in Figures 5.60 and 5.61.

Newspaper	***Guardian***	***Independent***	***Mail***	***Express***	**Total Neologism Usage**
Number of Neologism Uses	1057	650	1078	264	3049

Table 5.19: DDEB1+2 neologism uses across newspapers

Newspaper	***Guardian***	***Independent***	***Mail***	***Express***	**Total Neologism Usage**
Number of Neologism Uses	1147	836	744	164	2891

Table 5.20: DDEB3 neologism uses across newspapers

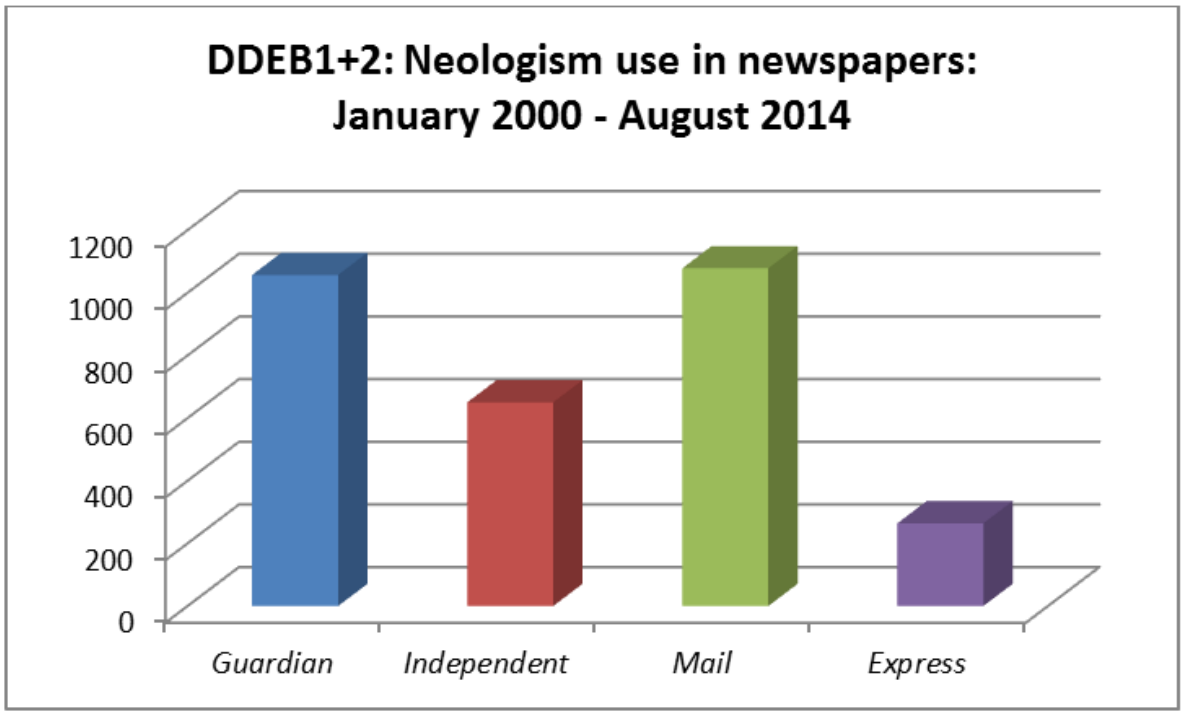


Figure 5.60: DDEB1+2 neologism use across newspapers

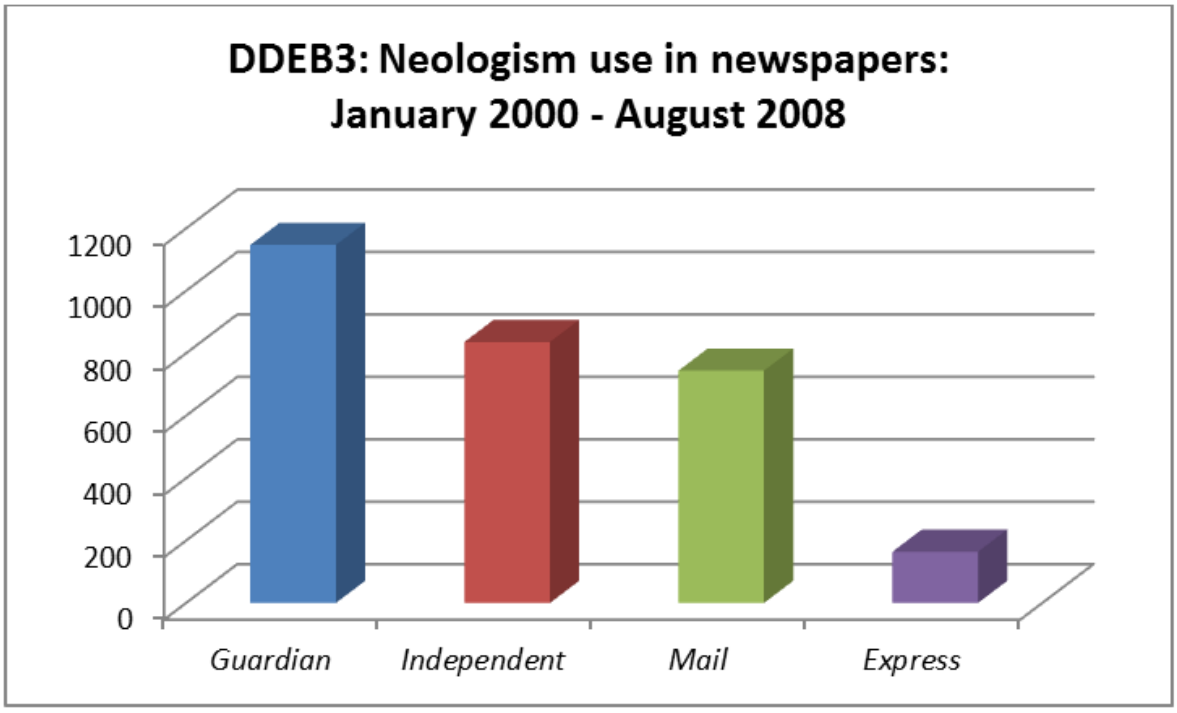


Figure 5.61: DDEB3 neologism uses across newspapers

Thus, in response to Research Question 2, one thing we can discover is that across the entirety of the neologic life-cycle in the context of this study (all three DDEBs), it is *The Guardian* which is most open to the use of neologisms. When we examine the study split into the two date ranges (2000-2008 and 2008-2014), again it is *The Guardian* which features neologisms most freely (DDEB3, 2008-2014). This is perhaps why studies of linguistics in newspapers tend to choose *The Guardian* as one of their key data sources (see for example Renouf (2007) and (2013) and Fischer (19980.

Monthly readership figures (across all platforms), according to newspaper marketing agency Newsworks (2015) are[147]:

- The *Mail* 30.6million
- *The Guardian* 27.6million
- The *Independent* 21.1million
- The *Express* 14.7million

[147] Based on July 2015-July 2016

This is excellent news for the development and spread of the neologisms, since they appear most often in the two newspapers with the highest monthly readership figures. It is interesting though that as neologisms become better established, they also become more prevalent in a newspaper with a slightly lower circulation, moving from the *Mail* (DDEB1+2) to *The Guardian* (DDEB3) as they progress through the neologic life-cycle.

As was outlined in Table 5.1, the 3,049 instances of the selected neologisms in DDEB1+2 were contained within 1,947 newspaper articles, while the 2,891 instances in DDEB3 appeared in 1,926 articles. Thus the entire set of 5,940 neologism occurrences was contained in 3,873 newspaper articles. Each article contained an average of 1.5 neologisms, but the majority contained a single instance. The highest number of instances in the same article was 18, for 'cyber-bullying'/'cyber bullying'.

In practice, this results in *The Guardian* producing 26 more articles containing neologisms than the *Mail* in DDEB1+2, yet the *Mail* carrying 21 more uses of DDEB1+2 neologisms than *The Guardian*. In DDEB3, *The Guardian* produces 123 more articles containing neologisms than the *Independent,* and 311 more neologism uses. This discrepancy is likely explained by the length of the articles themselves. One would expect that the longer the article, the more likely the possibility of multiple neologisms, with either the same word being repeated, or occasionally, more than one neologism in the same article.

When the average number of tokens per article was analysed (using Sketch Engine calculations of the running words in each file to ascertain the number of tokens in the text) it was found that, as expected, the articles in *The Express* were consistently shorter than all of the other newspapers, tying in with its consistently fewer neologisms. In DDEB1+2, the longest articles according to the average number of tokens were in the *Mail*. In DDEB3 they were in *The Guardian*. This helps to explain the findings regarding numbers of neologism uses in the two datasets; *The Guardian* achieves greater numbers of neologism uses across the study as a whole, as well as in DDEB3 specifically, simply because of the length of its articles and the average number of tokens per article. Returning to Research Question 2 then, we can see that within the neologic life-cycle,

the key element guiding levels of neologism use in individual newspapers is standard article length.

> **Research Question 2 –** *What can be discovered about the 'neologic life-cycle' of selected neologisms in UK national newspapers between 2000 and 2014?*

The central element for gathering and using all of the information above was having access to the publication date of each of the articles containing the neologisms under study. Without this key contextual component it would not have been possible to track the development of the new words over the 14 year period, and there would have been no way of discovering or presenting information on the differences in newspaper appearances between neologisms in different phases of their development.

> **Research Question 3 –** *In the context of data collection for context-rich, genre-specific web-based corpora, is the proposed new manual methodology more or less appropriate and effective in tracking neologism use and behaviour than the automated methods of the kind used by the* NeoCrawler*?*

In considering Research Question 3 above, and the issue of context-rich, genre-specific web-based corpora, we must bear in mind that, as discussed in 3.7.1, collecting dates for a web-based corpus is notoriously difficult even for automated systems (such as the *NeoCrawler*). Indeed there appears to be only one mechanism built into the web which offers any opportunity for dating such texts, and this is less than reliable (Kehoe 2006: 297-8). There are, from my personal experience, simply too many variables in how a date is presented, and too many points of confusion (such as month names being used as personal names (April, May, June for example)) to be able to collect this information automatically.

Unlike the *NeoCrawler*, the new methodology devised and tested here can avoid skewing results by using manual methods, such as the 'pre-screening' and 'advance exploration' of websites to exclude neologisms which appear only in links to other webpages, or in reader comments rather than in the newspaper article itself, and to therefore only select the articles actually required.

Thus it seems that in answer to Research Question 3, manual methods are **more** appropriate and effective for tracking neologism use and behaviour than automated methods of the kind used by the *NeoCrawler.*

5.4.2.1 Newspaper Neologism Usage and Emerging Dictionary Entries

In this section, in considering Research Question 2, I explore the relationship between newspapers and the most recent entrants into the neologic life-cycle, the newest of the dictionary entries, DDEB2 or those having as yet appeared only in *Wiktionary* or in a single expert-produced dictionary between September 2008 and August 2014.

Analysis of the newspaper usage of neologisms discussed in 5.4.2 reveals that while most research into neologisms and newspapers focuses on *The Guardian*, it is actually the *Independent* which has the strongest relationship with the newest of neologisms. More than any other newspaper, it features new words which have as yet entered only a single dictionary. It also appears to begin using these new words sooner than its competitors. In answering this Research Question then, it is clear that the most recent stage of the neologic life-cycle has more impact on and receives more feedback from the *Independent* newspaper than any other.

Once again, contextual information was crucial to gathering and tracking the data necessary to draw these conclusions. Publication dates, dates of dictionary inclusion, article labelling and the appearance of neologisms outside the parameters of this study (that is, pre-2000) all allow for a much more nuanced exploration of new words in newspapers than can be achieved through automated means.

This will prove useful for many genre-specific corpus linguistics projects, and, in my view provides a clear answer to Research Question 3:

> *In the context of data collection for context-rich, genre-specific web-based corpora, is the proposed new manual methodology more or less appropriate and effective in tracking neologism use and behaviour than the automated methods of the kind used by the* NeoCrawler*?*

That answer is that is more appropriate and effective in all areas.

Considering these issues in more detail, the position of the *Independent* in relation to new word appearances is an interesting one, since it actually appears to serve as a kind of bridge for new words transitioning between the two most recent stages of the neologic life-cycle. Not only does it include the second highest number of neologisms from DDEB1 (words not yet included in a dictionary), it is also almost always the newspaper to feature the most new words in DDEB2, which have as yet only entered *Wiktionary*.

When we examine the list of neologisms and the dictionaries they have entered, in Table 5.21 we can see that five words still appear only in *Wiktionary*: 'floordrobe', 'gendercide', 'globesity', 'superphone', 'wellderly'. ('Wellderly' is a slight anomaly, in that it falls under DDEB3, however it behaves more like a neologism appearing more recently in the neologic life-cycle, since it was actually accepted into *Wiktionary* on 27 August 2008, missed inclusion in DDEB2 by a mere four days.)

Dataset	Neologism	*Oxford English Dictionary*	*Oxford Dictionary of English*	*Oxford Dictionaries online*	*Merriam-Webster dictionary*	*Wiktionary*
DDEB3	acedia (n)	Y	Y	Y	Y	Y
DDEB2	bankster (n)	N	N	Y	N	Y
DDEB3	bogof* (n)	Y	Y	Y	N	Y
DDEB1	buzz marketing (n)	N/A	N/A	N/A	N/A	N/A
DDEB1	cold peace (n)	N/A	N/A	N/A	N/A	N/A
DDEB3	conurbation (n)	Y	Y	Y	Y	Y
DDEB2	cyberbullying* (n)	Y	N	Y	Y	N
DDEB2	cyberchondriac (n)	Y	N	Y	N	Y
DDEB2	diabesity (n)	N	N	Y	N	Y
DDEB3	earworm (n)	N	Y	Y	Y	Y
DDEB3	e-tailer (n)	Y	Y	Y	Y	Y
DDEB3	e-waste (n)	N	Y	Y	Y	Y
DDEB2	floordrobe (n)	N	N	N	N	Y
DDEB3	frenemy (n)	Y	Y	Y	Y	Y
DDEB2	gendercide (n)	N	N	N	N	Y
DDEB2	globesity (n)	N	N	N	N	Y
DDEB3	greenwashing* (n)	Y	Y	Y	Y	N
DDEB3	hubristic (adj)	Y	Y	Y	Y	Y
DDEB2	hyperlocal* (adj)	N	N	Y	N	Y
DDEB1	newer markets (n)	N/A	N/A	N/A	N/A	N/A
DDEB1	open education (n)	N/A	N/A	N/A	N/A	N/A
DDEB1	predatory lending (n)	N/A	N/A	N/A	N/A	N/A
DDEB3	promissory note (n)	Y	Y	Y	Y	Y
DDEB2	rewilding* (n)	Y	N	N	N	N
DDEB1	round pound (n)	N/A	N/A	N/A	N/A	N/A
DDEB1	sodcasting (v)	N/A	N/A	N/A	N/A	N/A
DDEB2	sovereign debt (n)	N	N	Y	N	Y
DDEB2	superphone* (n)	N	N	N	N	Y
DDEB1	tablet computing (n)	N/A	N/A	N/A	N/A	N/A
DDEB3	tenebrous (adj)	Y	Y	Y	Y	Y
DDEB3	upskill (v)	Y	Y	Y	N	Y
DDEB3	warrantless (adj)	Y	Y	Y	N	Y
DDEB3	waterboarding* (n)	Y	Y	Y	Y	Y
DDEB3	wellderly (n)	N	N	N	N	Y

Table 5.21: Neologism inclusion in dictionaries

*includes spelling variants
DDEB1; DDEB2; DDEB3

When we examine the spread of newspaper appearances for these five words (Table 5.22), we see that in all bar one case ('superphone') it is the *Independent* which uses them most.

Neologism	*Guardian*	*Independent*	*Mail*	*Express*
floordrobe	2	3	1	0
gendercide	4	8	7	0
globesity	1	6	3	0
superphone	2	5	17	5
wellderly	1	12	3	3

Table 5.22: DDEB2 single-dictionary neologisms appearing in the *Independent*

This, coupled with the fact that, in all bar one case in DDEB1+2 and one in DDEB3, the *Independent* is the newspaper featuring usage of neologisms before the start date for this research project (2000), suggests that, intentionally or otherwise, the *Independent* is most receptive to new words. We must bear in mind that, as explained above, the newspaper websites only include articles back to a certain date; after this, one must turn to separate archives, most of which are not available to individual researchers. This means that it is possible that there are earlier appearances of these neologisms in the other newspapers. However, the fact that almost all of the pre-2000 instances were in the *Independent* suggests that it is still the biggest user of neologisms. This is interesting given that previous studies have tended to focus most heavily on *The Guardian* when seeking to examine the relationship between new words and newspapers (see for example Renouf (2013) and Fischer (1998) (although Renouf's 2013 study did include both *The Guardian* and the *Independent*). Through exploration leading to answering Research Question 2 we can perhaps suggest that Renouf was right to include the latter, although whether this was the reason why remains unknown.

Gathering the data to make these observations was only possible due to the manual methods employed in creating the *NTON* database. By 'pre-screening' search results and conducting 'advance exploration' of associated websites, it was possible both to select articles of the correct date, but also to exclude inappropriate texts. Thus a neologism appearance in an article attributed to a press agency or paid for as part of a sponsorship agreement could be excluded as in the first case it was likely to appear in exactly the same form in all of the newspapers, and in the second, it was not written by a professional journalist. These kinds of nuances could not have been recognised by

automated systems like the *NeoCrawler*, and hence we can once again see, in response to Research Question 3, that the new manual methodology proposed here is **more** appropriate and effective in tracking neologisms use and behaviour than automated methods.

5.4.3 Factors Influencing Use and Development of Neologisms

Here, I discuss the influence of external factors on the way neologisms behave and develop in the media.

As mentioned in 5.2, it had originally been expected that increases in newspaper uses of neologisms would follow entry into dictionaries, particularly *Wiktionary*. Even after this had been shown not to be the case (by the lack of any consistent correlation between dictionary inclusion date and rising media usage, as shown in Tables 5.2 and 5.3), the underlying idea remained that there should be some relationship between a neologism's first dictionary entry and increasing levels of media usage. In fact, it was found that both tended to occur at around the same time, suggesting that some external influence was responsible. It seemed likely that this was actually some social or cultural factor, for example the increase in media use of 'waterboarding' seemed to coincide with an increase in international concern over interrogation techniques employed by the US.

When we re-examine the figures for each of the neologisms in the two datasets, we can see that, as Table 5.2 demonstrates, the unexpectedly high usage within DDEB1+2 is largely concentrated on just three words: 'cyberbullying', 'sovereign debt' and to a lesser extent, 'hyperlocal'. These account for 89.6% of all DDEB1+2 neologism use.

These three words entered dictionaries between 2009 and 2011, two of them ('hyperlocal' and 'sovereign debt') appearing in *Wiktionary*. Their entry dates coincide with increases in newspaper use of these words, which in turn fit the pattern of newspaper-to-neologism usage in DDEB2. Thus most of the 'sovereign debt' appearances are in *The Guardian* (457), the majority of those for 'cyberbullying' are in the *Mail* (658) and most of those for 'hyperlocal' are again in *The Guardian* (142).

Several of the DDEB3 entries into dictionaries similarly appeared to coincide with an uptake in media usage, for example 'hubristic' and 'waterboarding'.

It therefore seems that many neologisms enter dictionaries at the same time as they gain popularity in the media, but that neither one causes the other. Instead, both are responses to wider social, economic and cultural factors taking place around them. In almost all cases (where a specific date can be established) it is *Wiktionary* which these words first enter. This fact supports answers already presented for Research Question 1, by providing further evidence of *Wiktionary*'s responsiveness to neologisms as compared to that of expert-produced dictionaries.

Having barely been seen until 2006, use of the term 'waterboarding' increased three-fold in 2007, and twice more in 2008. It rose more than 60% more in 2010, before beginning to fall slightly. Consolidating its place in the lexicon, 'waterboarding' entered *Wiktionary* in 2007, was accepted into *OED* two years later and appeared in *ODE* in 2010. Usage remained high until 2014, when it returned to the 2007 level. It will be interesting to examine future usage levels to see if this decline continues, supporting Algeo's view of desuetude, the view that many neologisms eventually fall out of use (1993). In my opinion, the strong presence of 'waterboarding' over a period of eight years and the increased international focus on the United States' methods of intelligence gathering as part of the ongoing 'war on terror', make desuetude unlikely, at least in our lifetime, and it may be that instead, we will see 'waterboarding' reach a stable plateau of usage. In doing so, it would also leave the purview of the neologic life-cycle, having entered multiple dictionaries and achieved consistent year-to-year media usage; thus no longer being considered 'new' under the parameters of this study (see 1.1). We can thus, in answer to Research Question 2, learn that external social, economic and cultural factors can bring the neologic life-cycle to a premature end by effectively stabilising the position of a neologism within the lexicon.

The use of the term 'sovereign debt' in newspapers even more closely mirrors events in the wider world, rising and falling directly in line with the development, peak and gradual lessening of the financial crisis between 2009 and 2014, as Figure 5.62 shows.

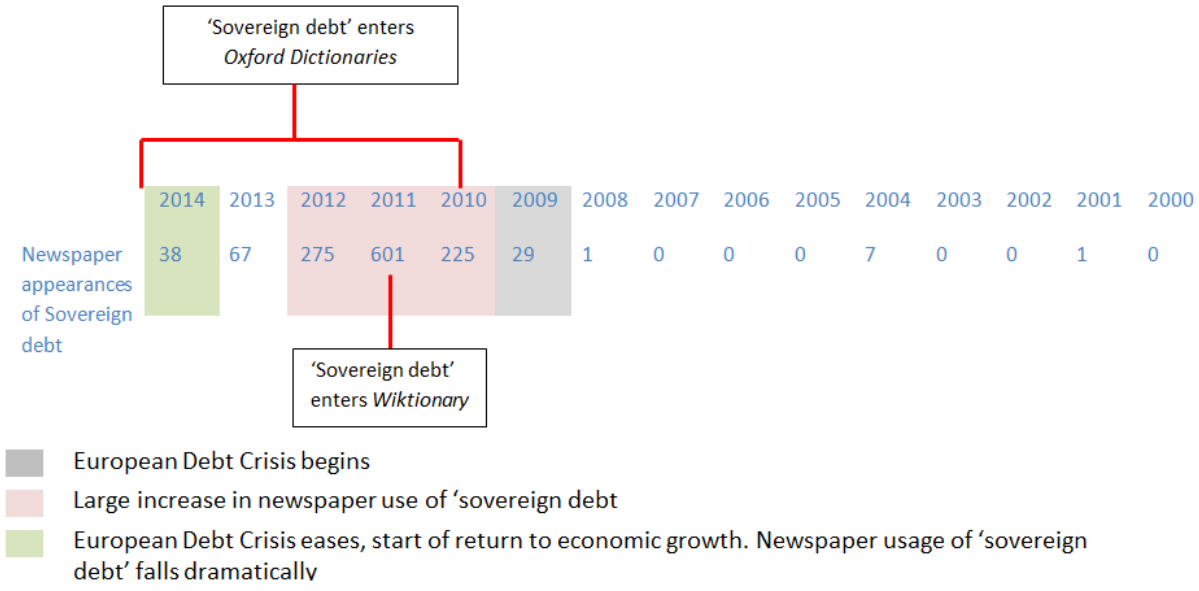


Figure 5.62: Life-cycle of 'sovereign debt', in relation to socio-economic factors

'Sovereign debt' entered *Wiktionary* in 2011 and *Oxford Dictionaries* online (*ODO*) sometime after 2010, although there is no way of knowing when. These dates correspond perfectly with the sudden increase in use of the term in newspapers, following the beginning of the European Debt Crisis. As this began to ease in 2014, the numbers of usage also began to fall.

In this instance, then, the external factors have the opposite effect to that for 'waterboarding', since they cause more pronounced inconsistencies in year-to-year usage of 'sovereign debt' in the media, thereby ensuring that the term remains firmly within the neologic life-cycle.

While 'cyberbullying' and 'hyperlocal' cannot be linked so directly with external factors, the rising instances of online harassment, coupled with several high-profile suicides[148] do appear to lend context to explain the increase in usage of the former, especially since the drop off in 2014 still leaves 168 instances of use in the four newspapers under study. It seems unlikely, given the ongoing problem of online bullying, that this number will fall further, although only time will tell. Changes in the nature of media during the past few years, meanwhile, can easily account for the rise in the use of the term

[148] See for example http://www.theguardian.com/society/2014/mar/10/self-harm-sites-cyberbullying-suicide-web

'hyperlocal'. For example there has been a move towards social media platforms within national newspapers and the growing legitimacy of the 'citizen journalist', defined by the *OED* as 'a non-professional journalist working outside traditional media channels; esp. a member of the public using the Internet and social media to publish news items or commentary'[149].

This does not account for the 26.92% drop in usage of 'hyperlocal' in 2014, however. It would be interesting to revisit these neologisms in these newspapers at a later date to establish whether this was just a temporary lapse or a genuine decline, and to seek an explanation in either case. It would also be interesting to do this in order to expand the current answers provided above for Research Question 2:

In summary then, it appears that the highest levels of neologism usage are most likely linked to non-linguistic factors such as the social, cultural and economic backdrop against which these new words are developing.

Of course just as social and economic factors influence the development of neologisms, so we also see these new words appearing more and more in popular culture. 'Earworm', for example, which entered *Wiktionary* in 2006, and *ODE/ODO* in 2010, was used in the *Doctor Who* episode *Under the Lake* (BBC), first broadcast on 3.10.15, and was central to an episode of the US comedy *The Big Bang Theory*, entitled *The Earworm Reverberation,* which aired in the UK on 24.12.15 (Channel 4). 'Frenemy', which entered *Wiktionary* in 2005, *OED* in 2008 and *ODE/ODO* in 2010, appeared in Stephen King's 2012 novel *11.22.63* (2012: 208).

As mentioned previously, the crucial element in gathering this data and being able to track the use of these neologisms with sufficient accuracy to be able to draw parallels both with dictionary entry and with the external factors around them, is date. Through the use of manual methods, the exact date of publication of all of these articles could be collected and used to position the appearance of these new words in the correct order, offering credible explanations for why certain words suddenly experience sometimes

[149] http://www.oed.com/view/Entry/33513?redirectedFrom=citizen+journalist#eid137461709

huge increases in usage, and at the same time are accepted into additional dictionaries. Having this information allows us to answer Research Question 3, that:

> *In the context of data collection for context-rich, genre-specific web-based corpora, the proposed new manual methodology is* ***more*** *appropriate and effective in tracking neologism use and behaviour than the automated methods of the kind used by the* NeoCrawler*?*

5.4.4 Conclusion – Media Tracking

In this section I summarise my findings from the Media Tracking project in light of the two Research Questions which I set out to answer:

> **Research Question 2 –** *What can be discovered about the 'neologic life-cycle' of selected neologisms in UK national newspapers between 2000 and 2014?*

> **Research Question 3 –** *In the context of data collection for context-rich, genre-specific web-based corpora, is the proposed new manual methodology more or less appropriate and effective in tracking neologism use and behaviour than the automated methods of the kind used by the* NeoCrawler*?*

As has been the case throughout the media tracking process, the crucial element in gathering this data has been the ability to accurately date every article. This has been possible only due to the highly detailed and contextualised data collected through the use of the manual methodology devised during the course of this project. While the basic search and tracking approach used by the *NeoCrawler* and myself were the same – conducting a search using Google – doing this manually produced more nuanced data than the *Neocrawler*'s preprogramed method could achieve. By examining and 'pre-screening' each Search Results Page, and then carrying out 'advance exploration' of the associated websites, it was possible to paint a picture of the use and development of neologisms over time.

In response to Research Question 2, meanwhile, it was discovered, that, for example, while *The Guardian* tends to be the newspaper of choice in studies conducted on neologisms in this context (see for example Fischer (1998) and Renouf (2013)), the

neologisms included there tend to be those which are better established, dating from earlier stages in the neologic life-cycle, belonging to Dictionary Date of Entry Batch 3 (DDEB3). The newspaper in which emerging neologisms are more likely to appear is the *Independent*. It includes the highest number of neologisms from the later stages of the life-cycle, DDEB2, those words having as yet only entered *Wiktionary*. However when adding all of the datasets together, *The Guardian* includes more neologism appearances across the entire neologic life-cycle covered by the 14-year period of this study, than any other newspaper.

It was also possible to examine the behaviour of words which had actually appeared in dictionaries for many years and yet which the *NeoCrawler* had erroneously collected as 'new'. I chose to retain these four words, to see how they might behave, since they had all entered *Wiktionary* in 2005 and 2006, which I took to be indication of 'newness', suggesting that perhaps these four words were entering a new phase in their own life-cycle. This did initially seem to be the case, since two out of these four words ('conurbation' and 'hubristic') saw increases in newspaper appearances around the time of their entry into *Wiktionary* (2005). However the other two 'reincarnated' words saw no such increase, and indeed all four have now gone into decline. These four words' continued fluctuation in usage, however, plus their inclusion as 'new' in both the *NeoCrawler* and *Wiktionary* (the latter during the period DDEB3) suggests that they cannot yet be considered to have completed the neologic life-cycle.

Crucial to achieving all of these results has been the date information collected as part of the piloting of this new methodology, as well as the additional elements of contextual information which ensured that only the most appropriate texts were selected for the *NTON* database. For example by excluding articles written by press agencies (which would be the same in every newspaper), articles paid for by external organisations (which were therefore not written by professional journalists) and articles written before the study period began (in 2000), the manual data collection methodology ensured that the information collected about neologism use and behaviour in UK national newspapers was as accurate and free from factors that might skew results as possible. While the *NeoCrawler* would be able to collect data from these newspapers by preprograming its Google search, it would not be able to narrow down

the search parameters in this way, and hence would produce broader but much less targeted results. This demonstrates, in answer to Research Question 3, that the proposed new manual methodology is **more** appropriate for tracking of neologism use and behaviour than are automated methods.

5.5 Conclusion

> **Research Question 1 –** *What can be learnt from this study about* Wiktionary's *responsiveness to neologisms and the level of detail and quality of definitions in its new word entries, when compared with expert-produced dictionaries?*

In response to Research Question 1, as discussed in 5.3.5, *Wiktionary* has been shown to provide more comprehensive entries for new words than expert-produced dictionaries, in terms of the number and the quality of the dictionary components which make up these entries, as well as the quality of the definitions within them. A key factor here is that *Wiktionary* is not constrained by the industry-standard dictionary components, but is free to include as much information as it chooses in its entries, in whatever style it deems fit. Contributors to the site are provided with templates to guide them, and the work each of them does on an entry is built upon by others, until a consensus is reached upon the optimum entry for that particular headword. In many cases this results in more detailed entries than are found in traditional dictionaries, whether 'corpus-based' or 'corpus-informed'.

A key element of these more detailed entries is their definition. Perhaps subconsciously using the same defining styles as expert-produced dictionaries, *Wiktionary's* collaborative contributors produce higher quality definitions which are clearer and more accessible than those of their competitors.

Wiktionary is also more responsive to neologisms than expert-produced dictionaries, not only due to the speed at which updates take place, but also from the work of contributors who, as was demonstrated through the many changes made to the entry for 'frenemy' can take a single, unstructured paragraph, to a full-blown, formatted entry

in just 24 hours. Every one of the changes made during this process resulted in the entire site being updated, something which takes months for expert-produced dictionaries.

> **Research Question 2 –** *What can be discovered about the 'neologic life-cycle' of selected neologisms in UK national newspapers between 2000 and 2014?*

With regard to Research Question 2 above, neologism appearances in UK newspapers have been tracked in order to compare usage and behaviour across their neologic life-cycle. This was done using the pilot of a new methodology for context-rich genre-specific corpus data collection, which was created as part of this project. This 'media tracking' showed how neologisms tend to move from one newspaper (the *Independent*) to another (*The Guardian*) as they move through the neologic life-cycle and become increasingly well-established. Across the entirety of this life-cycle, *The Guardian* includes more neologism appearances than any other newspaper, and the same is true in the earliest stage of the life-cycle, DDEB3 (dating back to January 2000). However in the most recent stages of the life-cycle (DDEB2, 2008-2014) neologisms which have only recently entered *Wiktionary* are much more likely to appear in the *Independent*.

The study also showed how usage of new words can suddenly increase due to external social, cultural or economic factors, and allowed for examination of words which appear to be entering a new phase in their own life-cycle. The former was shown to have the potential for *possibly* bringing a word to completion of the neologic life-cycle, by stabilising its number of year-to-year media uses (for example 'waterboarding'), or maintaining a word in its current position (for example 'sovereign debt'). In addition, it was discovered that the neologic life-cycle can still apply to 'reincarnated' words such as 'acedia', 'conurbation', 'hubristic' and 'tenebrous' even though they have been in use for many years, by virtue of the way they appear as 'new' for a time, fade away to the point of what Algeo would call 'desuetude' (1993) but then rise again, once more 'new'.

Research Question 3 – *In the context of data collection for context-rich, genre-specific web-based corpora, is the proposed new manual methodology more or less appropriate and effective in tracking neologism use and behaviour than the automated methods of the kind used by the* NeoCrawler*?*

In response to Research Question 3, it was found that the manual approach was more appropriate and more effective than the automated one, due to its ability to collect detailed contextual information. Crucial to this finding was date, in particular the publication date of newspaper articles containing neologisms. Without this information it would not have been possible to track the neologism usage so closely. In addition, the manual system allowed for inappropriate texts (for example where the neologism appeared in an advertisement rather than the article) to be excluded in advance, so that the texts which were collected, were those containing the right information in the right format. While the *NeoCrawler* could have collected data on newspaper articles containing neologisms, it would have taken much more of a 'broad brush' approach, and the results could have been contaminated (for example including articles not written by professional journalists) in a way which the manual methodology would be able to avoid.

Through the use of manual methods, the exact date of publication of all of these articles could be collected and used to position the appearance of these new words in the correct order, both within individual dictionaries, and across the media set at large. As discussed in 3.6.2, this level of nuanced data collection and tracking would have been impossible with the *NeoCrawler*'s style of automated searching and tracking. Although it collects data for particular date periods (the preceding week) it would have been unable to recognise all of the potential permutations of a 'date' that are to be found in newspaper articles, and hence would have either collected an incomplete set, or provided data that could not be placed into a useful order.

Chapter 6 Conclusion

6.1 Introduction

This lexicographical study set out to explore new words in the 'digital age', examining their representation in both expert-produced and collaborative dictionaries in order to compare degrees of comprehensiveness between them, as well as investigating their use and behaviour at different stages in the neologic life-cycle, in a real-word context: that of newspapers. In order to achieve the latter, a new methodology was created, designed for the collection of web-based data for context-rich genre-specific corpora. As well as being piloted by creating the *NTON* database (*Neologism Tracking in Online Newspapers*) this new methodology was compared with that of a similar but automated data collection system. The purpose was to assess which approach best enables the exploration of the use and behaviour of neologisms in the media.

The findings of the study showed that, despite this being a resolutely 'digital age', the new manual methods are better able to explore neologism use and behaviour in the media than automated ones, and this is due to contextual information. Researchers are able both to collect contextual information to provide more nuanced results, and to use this information to identify and 'pre-exclude' texts which could ultimately have skewed their results. The study's findings also showed that *Wiktionary* provides more comprehensive entries for neologisms entering its pages than do expert-produced dictionaries, and it is, on a wider scale, more responsive to new words than 'traditional' dictionaries. This is due to the flexibility *Wiktionary* enjoys as an independent, collaborative site.

6.2 Implications of Findings in the Wider Academic Context

The implications of these findings could be far-reaching. *Wiktionary*'s collaborative approach is, in my opinion, beginning to shift power over the language into the hands of the people. While it may only be a perception that lexicographers and academics have authority over language since they are the ones producing the reference sources that explain and codify it, it is a perception that I believe many people subconsciously share.

The introduction of dictionaries which mean that ordinary people can, if they choose, become involved in this process could begin to challenge this perception. This fact, coupled with the findings of this project suggesting that *Wiktionary* is more successful than expert-produced dictionaries in keeping up to date with the changes wrought in language by the addition of new words, may require the publishers of expert-produced dictionaries to begin considering major changes for the future. It is my belief, from the findings produced here, that in order to remain successful, traditional dictionary publishers may need to consider the possibility of joining forces with collaborative dictionaries in the next few years. At the very least, offering the kind of transparency over the dictionary-making process that *Wiktionary* provides through its Revision Histories and Discussion Forums would help to 'de-mystify' dictionaries produced by the lexicographers and academics who seem so remote from the ordinary man or woman on the street. Publishers may also need to relax their inclusion criteria, or perhaps introduce a 'pre-inclusion' section in their dictionaries (something akin to *Wiktionary*'s 'protologisms' page) where new words which have not quite achieved acceptance are included on a sort of trial basis. This would, of course, represent a significant change for historical dictionaries like the *OED*, however changes are already happening, with so many publishers pulling out of the print market, and instead focussing exclusively on their electronic offerings[150]. Further research would certainly be required before any changes were made, since one of the limitations of the current study was that only four expert-produced dictionaries and one collaborative one could reasonably be included. In addition, obtaining information on how different types of expert-produced dictionaries are created (that is, their relationship to corpora) was challenging here since, unsurprisingly, this information is not publicly available because it is considered commercially sensitive. In terms of collaborative dictionaries, whilst I believe that *Wiktionary* stands alone as the most comprehensive of those available, more data would be required on its competitors and on other non-traditional forms of dictionaries (such as portals) before any changes could be made.

[150]http://www.telegraph.co.uk/culture/books/booknews/7970391/Oxford-English-Dictionary-will-not-be-printed-again.html; http://www.theguardian.com/books/2012/nov/07/macmillan-dictionary-digital-finalprint.

An alternative approach for publishers might be to expand the semi-collaborative dictionary sections which several traditional publishers have been using in recent years (for example *Macmillan*'s *Open Dictionary*[151]). This is an area which particularly interests me, and represents the topic of the first piece of post-doctoral research I would like to conduct. Certainly it seems that Penta is correct when he argues: 'Now may be the time for dictionary makers to redefine themselves in the digital age, to plug into the collective and share its expertise of a truly ancient craft – and to allow the community to share its own sense of what a dictionary should be' (2011: 14).

There are implications also from the findings of the 'media tracking' and comparison of manual versus automated corpus data collection elements of my study. At first glance it may seem counterintuitive that in the modern digital age, the way to gather more detailed corpus data is to adopt manual methods. However I would argue that no matter how comprehensive and effective computers become, until artificial intelligence becomes a bona fide reality – until the Turing Test can truly be passed – there will always be value judgements associated with research that computer programs cannot make. (It is interesting to note that Renouf seeks to apply the Turing Test to neologisms undergoing automatic linguistic analysis (2006). Although she does so in a self-confessedly lighthearted manner, acknowledging factors that limit the possibility of an interrogator being able to distinguish between human and computer (the basis of the Test) (Ibid: 117, 120-1), she still, in my opinion closes with the inference that we are closer to being able to rely on computer judgement than I believe is the case. The new methodology created during this project presents a framework within which to make these human judgements, and to use them to generate more detailed and nuanced findings. The cost of that, of course, is that corpus linguists adopting this new methodology face the prospect of a sideways move to where digital and analogue methods combine. I believe the benefits will be worth the effort, however, since they offer the possibility of creating more nuanced corpora. Contextual information previously only available in the smallest of studies would become available to larger projects, and could lead to fascinating results. I believe the creation of this new methodology represents one of the key contributions of this project to academic study.

[151] http://www.macmillandictionary.com/open-dictionary/

This study has further served to make corpus linguistics researchers aware of an issue which may already have been affecting their research in negative ways; that of the Right to be Forgotten. Many researchers may not have been aware of the way in which webpages can now suddenly disappear from a site (and reappear on it) at the behest of individuals petitioning search engine providers like Google. While the pages still exist, and can be accessed through other means, this potentially presents a major problem for researchers, not only in terms of the implications for replicating studies, but also simply because when one returns to a site to check results or gather additional information at a later date, it may have completely disappeared, leading to sometimes last minute changes to the research findings.

Finally, of course, this study and its findings have begun to fill gaps in the academic research identified at the beginning of the project, for example on comparisons of different dictionary formats, methodologies for working with neologisms and, centrally, the relationship between lexicography and neology.

6.3 Looking to the Future

The issues raised in 6.2 of course have implications for future research in the fields of lexicography and neology, as well as for corpus linguistics as a tool for dictionary-makers. One such is the existence of an alternative methodology for carrying out web-based data collection for context-rich genre-specific corpora. It is hoped that this will prove useful in a wide variety of fields, enabling researchers to more effectively narrow down the texts they select, and produce the more nuanced results discussed above. Of course further testing of the new methodology will be required, and it will be interesting to see how it is used and developed through this process. One particularly interesting idea, in my view, would be to conduct a study similar to this one, but on transcripts from broadcast media; television and radio broadcasts continuing in the sphere of professional journalism, and podcasts and 'vlogs' (video logs) moving back into the field of social media. I would also like to conduct further work on the idea of the 'neologic life-cycle', working on a generalised definition applicable to wider

neologism studies, identifying and establishing clear 'stages' within this period, and even developing a 'core' neologic life-cycle which could then be augmented for individual studies through the use of a 'toolkit' of add-on elements.

Corpus linguists should also now, in my view, take extra care to cache and download webpages they hope to use in their research *as they go along*, in order to avoid the difficulties presented by the Right to be Forgotten legislation if they need to go back to a page, and find that it has disappeared.

In lexicography and neology, there is definite scope for a full-scale research project into *Wiktionary*'s Revision History and Discussion Forum functions, since it is these that are responsible for the levels of transparency which *Wiktionary* can offer its users, and which expert-produced dictionaries cannot hope to match. I believe such research could shed important light on how the future of the dictionary landscape might develop, and would enable studies, like this one, which bring together the academic fields of lexicography and neology, an area which, as shown in Chapter 2, has achieved so little attention.

Finally, it would be interesting to return to the Oxford University Press' *New Words Corpus*[152], started in 2012 and as yet unavailable to researchers. Already words in my 'not entered a dictionary' category have been included in some of the dictionaries I have investigated, for example 'sodcasting' was accepted into *Wiktionary* several months after data collection for this study ended, while 'bankster' and 'hyperlocal' similarly entered the *Oxford English Dictionary.* What new words might we find in the *New Words Corpus*, when (or if) it is made an open resource?

[152] https://en.oxforddictionaries.com/explore/oxford-new-words-corpus

References

Abel, A. and Meyer, CM. (2013) 'The Dynamics Outside the Paper: User Contributions to Online Dictionaries' *Proceedings of the eLex 2013 Conference,* 'Electronic Lexicography in the 21st Century: Thinking Outside the Paper'. Held 17-19 October 2013 in Tallinn, Estonia. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, 179-194. Available from <http://eki.ee/elex2013/proceedings/eLex2013_13_Abel+Meyer.pdf> [14 December 2013]

Algeo, J. (1980) 'Where Do All The New Words Come From?', *American Speech*, 55(4), 264-277

Algeo, J. (1993) 'Desuetude among New English Words'. *International Journal of Lexicography* 6(4), 281-293

Alpert, J. and Hajaj, N. (2008) cited in Fletcher, W.H. (2013) 'Corpus Analysis of the World Wide Web'. in *The Encyclopedia of Applied Linguistics*. ed. by Chapelle, C.A. [online] Blackwell Publishing. Available from <http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0254/full> [12 February 2013]

Atkins, BTS. and Rundell, M. (2008) *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press

Baayen, R.H. and Renouf, A.J. (1996) 'Chronicling the Times: Productive Lexical Innovations in an English Newspapers', in *Language*, 72(1) 69-96

Barrs, K. (2015) 'Catachrestic and Non-Catachrestic English Loanwords in the Japanese Language'. in Formato, F. and Hardie, A. ed. *Abstract Book of Corpus Linguistics 2015*, at Lancaster: UCREL. 372-374

Biber, D. (2008) 'Representativeness in Corpus Design' in *Practical Lexicography: A Reader*. ed. by Fontenelle, T. Oxford: Oxford University Press, 63-87

Bös, B. (2012) ’From 1760 to 1960: Diversification and Popularization’. in *News as Changing Texts.* ed. by Facchinetti, R. Brownlees, N., Bös, B. and Fries, U. Newcastle-upon-Tyne: Cambridge Scholars Publishing, 91-143

Brownlees, N. (2012) ‘The Beginnings of Periodical News (1620-1665)’. in *News as Changing Texts.* ed. by Facchinetti, R. Brownlees, N., Bös, B. and Fries, U. Newcastle-upon-Tyne: Cambridge Scholars Publishing, 5-48

Bryant, SL., Forte, L. and Bruckman, A. (2005). ‘Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia’, *GROUP '05 Proceedings of the 2005 ‘*International ACM SIGGROUP Conference on Supporting Group Work’, held 6-9 November 2005 on Sanibel Island, Florida, 1-10

Bryer, T. (2013) ‘Designing Social Media Strategies for Effective Citizen Engagement: A Case Example and Model’. *National Civic Review*, Spring, 43 – 50

Businessballs.com (2015) *Demographics Classifications* [online]. Available from <http://www.businessballs.com/demographicsclassifications.htm#nrs-social-grade-definitions-uk> [30 March 2015]

Carr, M. (1997) ‘Internet Dictionaries and Lexicography’. *International Journal of Lexicography* 10(3), 209-230

Carstairs-McCarthy, A. (2002) *An Introduction to English Morphology’*. Edinburgh: Edinburgh University Press

Chen, G. (2013) ‘The Impact of the Medium on OED1-3’, *OED Symposium 2013 Newsletter Issue 7.* Oxford: Oxford University Press, 4-6

Cohen, L., Manion L., and Morrison, K. (2000) *Research Methods in Education* Abingdon: RoutledgeFalmer

Cole, P. and Harcup. T. (2009) *Newspaper Journalism*. London: SAGE Publications

Cox, S. (2014) BBC Radio 4 ‘The Right to be Forgotten’, *The Report* [18 September 2014, 8pm]

Creese, S. (2015). 'In Search of Oblivion? How the 'Right to be Forgotten' could Undermine Web-Based Corpora'. *Procedia – Social and Behavioral Sciences. Current Work in Corpus Linguistics: Working with Traditionally-Conceived Corpora and Beyond. Selected Papers from the 7th International Conference on Corpus Linguistics (CIL2015).* (ed.) by Fuertes-Olivera, PA., Álvarez de la Fuente, E., Fernádez-Fuertes, R., Garcés García, P., López Arroyo, B., Niño Amo, M., Pizarro Sánchez, I., Sáez-Hidalgo, A., Satre-Ruano, M[a.] Angeles, and Velasco-Sacristán, M. Vol 198. 95-102. Available from http://ac.els-cdn.com/S1877042815044250/1-s2.0-S1877042815044250-main.pdf?_tid=c6382290-8c53-11e5-a3ac-00000aacb361&acdnat=1447672894_97d1451264973033dce57ef8c0385f87. [15 June 2016]

Danesi, M. (2016) *Language, Society and New Media. Sociolinguistics Today*. Abingdon: Routledge

Doctor Who (2015) BBC, 3 October 2015

Dörnyei, Z. (2007) *Research Methods in Applied Linguistics*. Oxford: Oxford University Press

Duffy, B. and Rowden, L. (2005) *You are what you Read? How Newspaper Readership is Related to Views* [online]. London/Edinburgh: Ipsos Mori. Available from <https://www.ipsos-mori.com/Assets/Docs/Publications/sri_you_are_what_you_read_042005.pdf>. [9 September 2016]. Cited on Businessballs.com (2015) *Demographics classifications* [online]. Available from <http://www.businessballs.com/demographicsclassifications.htm#nrs-social-grade-definitions-uk> [30 March 2015]

European Commission (2014) *Factsheet on the 'Right to be Forgotten' Ruling*, [online] available from <http://ec.europa.eu/justice/dataprotection/files/factsheets/factsheet_data_protection_en.pdf> [1 March 2015]

Evans, MP. (2007) 'Analysing Google Rankings through Search Engine Optimization Data', *Internet Research*, 17(1), 21-37

Facchinetti, R. (2012) 'News Writing from the 1960s to the Present Day'. in *News as Changing Texts.* ed. by Facchinetti, R. Brownlees, N., Bös, B. and Fries, U. Newcastle-upon-Tyne: Cambridge Scholars Publishing, 145-195

Ferraresi, A., Zanchetta, E., Baroni, M. and Bernardini, S. (2008) 'Introducing and Evaluating ukWaC, A Very Large Web-Derived Corpus of English' *Proceedings of the 4th Web as Corpus Workshop* (WAC-4) 'Can we beat Google?'. Held 1 June 2008 in Marrakech, Morocco

Fisher, T. (2014) 'What is an XML File?', *Lifewire* [online]. Available from https://www.lifewire.com/what-is-an-xml-file-2622560 [1 September 2015]

Fischer, R. (1998) *Lexical Change in Present-Day English*. *A Corpus-Based Study of the Motivation, Institutionalization, and Productivity of Creative Neologisms*. Tübingen: Gunter Narr Verlag

Fletcher, W.H. (2013) 'Corpus Analysis of the World Wide Web'. in *The Encyclopedia of Applied Linguistics*. ed. by Chapelle, C.A. [online] Blackwell Publishing. Available from <http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0254/full> [12 February 2013]

Flowerdew, J. (2013) *Discourse in English Language Education*. Abingdon: Routledge

Fortunato, S., Boguna, M., Flammini, A. and Menczer, F. (2006) 'How to Make the Top Ten: Approximating PageRank from In-degree'*, Paper presented at the 14th International World Wide Conference,* held May 22-26. in Edinburgh. Available from <http://xxx.lanl.go/PS_cache/cs/pdf/0511/0511016.pdf>

Fox, G. (1987) 'The Case for Examples' in *Looking Up. An Account of the COBUILD Project. in Lexical Computing*. ed. by Sinclair, JM. London: Collins, 137-149

Francl, M. (2011) 'Neolexia'. *Nature Chemistry*, 3, 417-418

Fries, U. (2012) 'Newspapers from 1665 to 1765'. in *News as Changing Texts.* ed. by Facchinetti, R. Brownlees, N., Bös, B. and Fries, U. Newcastle-upon-Tyne: Cambridge Scholars Publishing, 49-89

Ginzburg, RS., Khidekel, SS., Knyazeva, GY. and Sankin, AA. (1979) *A Course in Modern English Lexicology*. Moscow: Vyssaja Skola

Greenslade (2012) *The Guardian – Message to Advertisers – Farewell Newspapers, Hello Newsbrands* [online] 22 May 2012. Available from <https://www.theguardian.com/media/greenslade/2012/may/21/national-newspapers-advertising> [30 July 2016]

Grefenstette, G. (2002) 'The WWW as a Resource for Lexicography'. in *Lexicography and Natural Language Processing*. A Festschrift in Honour of BTS Atkins. ed. by Corréard, M-H. UK: Euralex 2002. 199-215

Hanks, P. (2012) 'Corpus Evidence and Electronic Lexicography'. in *Electronic Lexicography*. ed. by Granger, S. and Paquot, M. Oxford: Oxford University Press, 57-82

Hanks, P. (2013) 'Creatively Exploiting Linguistic Norms'. in *Applications of Cognitive Linguistics (ACL): Creativity and the Agile Mind: A Multi-Disciplinary Study of a Multi-faceted Phenomenon.* ed. by Veale,T., Feyaerts, K. and Forceville, CJ. Berlin/Boston: De Gruyter Mouton. 119-138

Hemsley, J. and Mason, RM. (2012) 'The Nature of Knowledge in the Social Media Age: Implications for Knowledge Management Models'*, Proceedings of the 45th Hawaii International Conference* 'System Sciences'. Held 4-7 January 2012, in Maui, Hawaii. 3928-3937

Hundt, M., Nesselhauf, N. and Biewer, C. (2007) 'Corpus Linguistics and the Web', *Language and Computers – Studies in Practical Linguistics*, 59, 8-12

Hunston, S. (2002) *Corpora in Applied Linguistics*, Cambridge: Cambridge Applied Linguistics Series, Cambridge University Press

Janssen, M. (2013). 'Lexical Gaps', in *The Encyclopedia of Applied Linguistics*. ed by Chapelle, C.A. [online] Blackwell Publishing. Available from: <http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0693/full> [21 November 2013].

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. and Suchomel, V.. 2013. 'The TenTen Corpus Family'. Paper presented at the *7th International Corpus Linguistics Conference*, Lancaster, July 2013. 125-127

Katamba, F. (1994) *English Words*. London: Routledge

Kehoe, A. and Renouf, A.J. (2002) 'WebCorp: Applying the Web to Linguistics and Linguistics to the Web'. *WWW 2002 The Eleventh World Wide Web Conference*. Held 7-11 May 2002, Honolulu, Hawaii

Kehoe, A. (2006) 'Diachronic Linguistic Analysis On The Web With WebCorp'. in *The Changing Face of Corpus Linguistics*. ed. by Renouf A. and Kehoe, A. Amsterdam: Rodopi (2006). 297-307

Kennedy, G. (1998) *An Introduction to Corpus Linguistics*. Harlow: Addison Wesley Longman

Kerremans, D. (2012) A Web of Words. On the Conventionalisation of English Neologisms Unpublished PhD thesis. Munich: Ludwig-Maximilians University

Kerremans, D. (2015) *A Web of Words. A Corpus-Based Study of the Conventionalization Process of English Neologisms*. Frankfurt: Peter Lang GmbH

Kerremans, D., Stegmayr, S. and Schmid, H-J. (2012) 'The NeoCrawler: Identifying and Retrieving Neologisms from the Internet And Monitoring on-Going Change'. in *Current Methods in Historical Semantics*. ed. by Allan, K. and Robinson, JA. Berlin/Boston: De Gruyter Mouton 59-96

Kilgarriff, A. (2013) 'Using Corpora as Data Sources for Dictionaries'. in *The Bloomsbury Companion to Lexicography*. ed. by Jackson, H. London: Bloomsbury. 77-96. Chapter downloads from Sketch Engine Bibliography as *Using Corpora [and the*

web] as Data Sources for Dictionaries. See https://www.sketchengine.co.uk/bibliography-of-sketch-engine/#toggle-id-2 [4 May 2016]

Kilgarriff, A. and Grefenstette, G. (2008) 'Introduction to the Special Issue on the Web as Corpus'. in *Practical Lexicography: A Reader*. ed. by Fontenelle, T. Oxford: Oxford University Press, 89-101

Kilgarriff, A. Husák, M. McAdam, K., Rundell, M. and Rychlý, P. (2008). 'GDEX: Automatically Finding Good Dictionary Examples in a Corpus'. in Bernal, E. and DeCesaris, J. (eds.) *Proceedings of the 13th EURALEX International Congress.* Held 15-19 July 2008, in Barcelona, Spain. 425–432

Kilgarriff, A. and Kosem, I. (2012) 'Corpus Tools for Lexicographers'. in *Electronic Lexicography*. ed. by Granger, S. and Paquot, M. Oxford: Oxford University Press, 31-55

King, S. (2012) *11.22.63.* London: Hodder and Stoughton

Køhler Simonsen, H. (2005) 'User Involvement in Corporate LSP Intranet Lexicography' in Gottlieb, H., Mogensen, JE. and Zettersten, A. (eds.) *Proceedings of the Eleventh International Symposium on Lexicography, '*Symposium on Lexicography XI'. held 2-4 May 2002, at the University of Copenhagen. Tübingen: Niemeyer, 489-510

Lee, D. (2014b) 'Google Reinstates 'Forgotten' Links after Pressure', *BBC News Technology*, [online] available from: <http://www.bbc.co.uk/news/technology-28157607> [28 February 2015]

Lehrer, A. (2003) 'Understanding Trendy Neologisms'. *Rivista di Linguistica,* 15(2), 371-384

Laufer, B. (2008) 'Corpus-based versus Lexicographer Examples in Comprehension and Production of New Words'. in *Practical Lexicography*. ed. by Fontenelle, T. Oxford: Oxford University Press. 213-218

Lew, R. (2011) 'Online Dictionaries of English'. in *e-Lexicography. The Internet, Digital Initiatives and Lexicography.* ed. by Fuertes-Olivera, PA., and Bergenholtz, H. London and New York: Bloomsbury Academic, 230-250

Lew, R. (2012) 'How can we make Electronic Dictionaries more Effective?'. in *Electronic Lexicography*. ed. by Granger, S. and Paquot, M. Oxford: Oxford University Press, 343-361

Lew, R. (2013) 'User-Generated Content (UGC) in English Online Dictionaries'. in *Ihr Beitrag Bitte! – Ker Nutzerbeitrag im Wörterbuchprozess (OPAL – Online Publizierte Arbeiten zur Linguistik),* ed. by Abel, A. and Klosa, A. Mannheim: Institut für Deutsche Sprache, 9-30

Lüdeling, A., Evert, S. and Baroni, M. (2007) 'Using Web Data for Linguistic Purposes', *Language and Computers – Studies in Practical Linguistics*, 59, 14-31

Leuf, B. and Cunningham, W. (2001). *The Wiki Way: Quick Collaboration on the Web*, USA: Addison-Wesley.

Matthews, PH. (2003) *Linguistics. A Very Short Introduction*. New York: Oxford University Press

McIntosh, N. (2015) 'List of BBC Web Pages which have been Removed from Google's Search Results' *BBC Internet Blog*, 25.06.15 [online] Available from <http://www.bbc.co.uk/blogs/internet/entries/1d765aa8-600b-4f32-b110-d02fbf7fd379> [19 December 2015]

Melchior, L. (2012) 'Halbkollaborativität und Online-Lexikographie. Ansätze und Überlegungen zu Wörterbuchredaktion und Wörterbuchforschung am Beispiel LEO Deutsche-Italienisch' in *Lexicographica* 28. ed. by Gouws, R.H., Heid, U., Schierholz, St.J., Schweickard, W. and Wiegand, H.E. Berlin/New York: de Gruyter. 337-372

Meyer, C. and Gurevych, I. (2010) 'How Web Communities Analyse Human Language: Word Senses in Wiktionary'. *Proceedings of the Second Web Science Conference*, held 26-27 April 2010 in Raleigh, NC, USA

Meyer, C. and Gurevych, I. (2012) 'Wiktionary: a New Rival for Expert-Built Lexicons? Exploring the Possibilities of Collaborative Lexicography'. in *Electronic Lexicography*. ed. by Granger, S. and Paquot, M. Oxford: Oxford University Press, 259-291

Minkova, D, and Stockwell, R. (2009) *English Words. History and Structure*. Cambridge: Cambridge University Press

Mitchell, RL. (2008) 'My Word: Why Google is in the Dictionary but AJAX isn't'. *ComputerWorld*, 27 October 2008, 32-34

Moon, R. (2008) 'Lexicography and Lexical Creativity'. *Lexikos* 18 (AFRILEX-reeks/series 18), 131-153

Moon, R. (2009) 'The Cobuild Project'. in *The Oxford History of English Lexicography, Volume 2 Specialized Dictionaries*. ed. by Cowie, AP. Oxford: Oxford University Press 436-457

Mugglestone, L. (2011) *Dictionaries: A Short Introduction.* Oxford: Oxford University Press

Nesi, H. (2009) 'Dictionaries in Electronic Form'. in *The Oxford History of English Lexicography, Volume 2 Specialized Dictionaries*. ed. by Cowie, AP. Oxford: Oxford University Press. 458-478

Nesi, H. (2012) 'Alternative e-Dictionaries: Uncovering Dark Practices'. in *Electronic Lexicography*. ed. by Granger, S. and Paquot, M. Oxford: Oxford University Press, 363-377

Neuman, Y., Nave, O. and Dolev, E. (2010) 'Buzzwords on Their Way to a Tipping Point: a View from the Blogosphere'. *Complexity*, 16(4), 58-68

Penta, D.J. (2011) 'The Wiki-fication of the Dictionary: Defining Lexicography in the Digital Age', *Media in Transition 7 Conference*, held 13 May 2011 at Massachusetts Institute of Technology, Cambridge, MA, USA

Preston, R. (2014) 'Why Has Google Cast Me into Oblivion?' *BBC Business News*, 2 July 2014. Available from: <http://www.bbc.co.uk/news/business-28130581> [28 February 2015]

Pringle, G., Allison, L. and Dowe, DL. (1998) 'What is a Tall Poppy among Web Pages?' *Proceedings of the 7th International Word Wide Web Conference, held 14-18 April, in* Brisbane, Australia. 369-77, available from www.csse.monash.edu.au/~lloyd/tilde/InterNet/Search/1998/WWW7.html

Renouf, A. (1987) 'Corpus Development'. in *Looking Up. An Account of the COBUILD Project in Lexical Computing*. ed. by Sinclair, JM. London: Collins, 1-40

Renouf, A. (2003). 'WebCorp: Providing A Renewable Data Source For Corpus Linguists', in *Extending The Scope Of Corpus-Based Research: New Applications, New Challenges.* ed. by Granger, S. and Petch-Tyson, S. Amsterdam and New York: Rodopi. 39-58

Renouf, A. (2006) 'The Turing Test Applied to Automatic Linguistic Analysis'. in *Linguists (don't) Only Talk About It: Essays in Honour of Bernhard Kettemann*. ed. by Fill, A., Marko, G., Newby, D. and Penz, H. Tübingen: Stauffenburg. 117-122

Renouf, A. (2007) 'Tracing Lexical Productivity and Creativity in the British Media: "The Chavs and The Chav-Nots". in *Lexical Creativity. Texts and Contexts*. ed. by Munat, J. Amsterdam: John Benjamins Publishing. 61-89

Renouf. A. (2013) 'A Finer Definition of Neology in English. The Life-Cycle of a Word' in *Corpus Perspectives on Patterns of Lexis*. ed. by Hasselgåard, H., Ebeling, J. and Oksefjell Ebeling, S. Amsterdam: John Benjamins Publishing, 177-208

Renouf, A. and Kehoe, A. (2013) ‘Filling the gaps Using the WebCorp Linguist’s Search Engine to Supplement Existing Text Resources’, *International Journal of Corpus Linguistics* 18(2). 167-198

Renouf, A., Kehoe. A and Banerjee, J. (2005) 'The WebCorp Search Engine: a Holistic Approach to Web Text Search', in *Electronic Proceedings of CL2005*, University of Birmingham.

Renouf, A., Kehoe, A. and Banerjee, J. (2007) ‘Webcorp: An Integrated System for Web Text Search’. in *Corpus Linguistics and the Web*. ed. by Hundt, M., Nesselhauf, N. and Biewer, C. Amsterdam: Rodopi, 47-67

Reah, D. (1998) *The Language of Newspapers*. Abingdon: Routledge

Robson, C. (2002) *Real World Research*. Oxford: Blackwell Publishers

Sarantakos, S. (1998) *Social Research*. London: Macmillan. Cited in Robson, C. (2002) *Real World Research*. Oxford: Blackwell Publishers

Sinclair, J. (1987) ‘The Nature of the Evidence’. in *Looking Up: An Account of the COBUILD Project in Lexical Computing*. ed. by Sinclair, JM. London: Collins, 150-159

Sinclair, J. (2004) ‘Corpus and Text – Basic Principles’. in *Developing Linguistic Corpora: A Guide to Good Practice.* ed. by Wynne, M. [online] Oxford: Oxbow Books. Accessed at: <www.ahds.ac.uk/linguistic-corpora/>. Accessed 21 February 2013. 5-24

Stevenson, A. (2010) *Oxford Dictionary of English*. Oxford: Oxford University Press

Storrer, A. (2010). ‘Deutsche Internet-Wörterbücher: Ein Überblick’. in *Lexicographica 27.* ed. by Gouws, RH., Heid, U., Schierholz, St.J., Schweickard, W. and Wiegand, HE. Berlin/New York: de Gruyter, 155–164.

Svensén, B. (2009) *A Handbook of Lexicography*. Cambridge: Cambridge University Press

The Big Bang Theory (2015) Channel 4, 24 December 2015

Tognini-Bonelli (2001) *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins Publishing Company

Trinity Mirror (2016) *New Day*. 29 February-6 May 2016

Weiner, E. (2009) 'The Electronic OED: The Computerization of a Historical Dictionary'. in *The Oxford History of English Lexicography, Volume 1 General-Purpose Dictionaries*. ed. by Cowie, AP. Oxford: Oxford University Press 378-409

Williams, R. (2015) 'Telegraph Stories affected by EU 'Right to be Forgotten', *The Telegraph* [online] 3 September 2015. Available from <http://www.telegraph.co.uk/technology/google/11036257/Telegraph-stories-affected-by-EU-right-to-be-forgotten.html> [19 December 2015]

Personal Communications

Bennett, A. (2014) *Building Web-Based Corpora* [conversation at BALEAP PIM] with S. Creese [21 June 2014]

Coventry University IT Services (2014) Express Search Results [informal conversation] with S. Creese [17 November 2014]

Kerremans, D. (2013) NeoCrawler – *Next Steps* [informal conversation] with S. Creese [4 December 2013]

Kilgariff A. (2014) *Using WebBootCaT* [email] to S. Creese [10 August 2014]

Zhang, Y. (2014) *Web-Based Corpus Building* [meeting] with S. Creese [22 May 2014]

Zhang, Y. (2014) *Web-Based Corpus Building* [informal conversation] with S. Creese [20 June 2014]

Suchomel. S. (2015) *WebBootCaT Issues in Sketch Engine* [email] to S. Creese [8 July 2015]

Whitelock, P. (2016) *Using the Oxford English Corpus* [email] to S. Creese [10 May 2016]

Mass Communication

Pezik, P. (2016), '*Monco - An Experimental Monitor Corpus Search Engine'* Email to Corpora Digest mailing list: corpora@uib.no*,* [January 11 2016]
http://monitorcorpus.com [19 February 2016]

Websites

Associated Newspapers (2014) *Mail Online* [online] available from <http://www.dailymail.co.uk/home/index.html> [17 December 2014]

EnerG (n.d.) - English Neologisms Research Group, Ludwig maximilians university munich – NeoCrawler [online] available from <http://www.neocrawler.de/crawler/html/> [1 January 2016]

Guardian News and Media (2014). *The Guardian* [online] available from <http://www.guardian.co.uk/> [17 December 2014]

Google News (2016) *Google News* [online] available from <https://news.google.co.uk/> [19 July 2016]

Independent.co.uk (2014) *The Independent* [online] available from <http://www.independent.co.uk/> [17 December 2014]

Internet for Lawyers (n.d.) *Google Kills Blogsearch - But Here's How You Can Force Google to Display it* [online] available from <http://www.netforlawyers.com/content/google-kills-blog-search-engine-109> [19 July 2016]

Internet for Lawyers (2016) *Google Kills Blogsearch (Again) - But This Time They REALLY Mean It* [online] available from

<http://www.netforlawyers.com/content/google-kills-blog-search-engine-109> [19 July 2016]

Ludwig-Maximilians Universität München (n.d.) *NeoCrawler* [online] available from <http://www.neocrawler.de/crawler/html/> [1 February 2014]

Merriam-Webster (2015a) *About Us* [online] available from <http://www.*Merriam-Webster*.com/about-us> accessed 22 August 2016

Merriam-Webster (2015b) *About Us – FAQs* [online] available from <http://www.Merriam-Webster.com/about-us/faq> accessed 22 August 2016

Merriam-Webster (2015c) *Help – FAQ - How does a word get into a Merriam-Webster dictionary?* [online] available from <http://www.merriam-webster.com/help/faq-words-into-dictionary> accessed 22 August 2016

News Group Newspapers (n.d.) *The Sun* [online] available from <http://www.thesun.co.uk/sol/homepage/> [18 December 2014]

Newsworks (2015) Facts and Figures [online] available from <http://www.newsworks.org.uk/Facts-Figures> [9 November 2016]

Newsworks (2016) *Newsworks – NRS PADD: 47 Million Read Newsbrands Monthly* [online] available from <http://www.newsworks.org.uk/News-and-Opinion/nrs-padd-90-of-gb-reads-newsbrands-every-month-> 2 June 2016 [30 July 2016]

Northern and Shell Media Productions (2014). *Express* [online] available from <http://www.express.co.uk/> [16 December 2014]

Nuance (2014) *Dragon Naturally Speaking Version 12.0* [online] available from <http://www.nuance.co.uk/for-business/by-product/dragon/dragon-for-the-pc/dragon-professional/index.htm> [18 December 2014]

Oxford University Press (2016a) *Catachresis* [online] available from <http://www.oed.com/view/Entry/28665?redirectedFrom=catachresis#eid> [21 August 2016]

Oxford University Press (2016c) *Reading Programme* [online] available from <http://public.oed.com/history-of-the-oed/reading-programme/> [21 August 2016]

Oxford University Press (2016d) *What are the Main Differences Between the OED and Oxford Dictionaries?* [online] available from <http://www.oxforddictionaries.com/words/what-are-the-main-differences-between-the-oed-and-odo> [accessed 21 August 2016]

Oxford University Press (2016e) 'The Corpus Reaches New Heights', *What is a Corpus* [online] available from <http://www.oxforddictionaries.com/words/what-is-a-corpus> [1 September 2016]

Oxford University Press (2016f) *The Oxford New Words Corpus (New Monitor Corpus)* [online] available from <http://www.oxforddictionaries.com/words/oxford-new-words-corpus> [21 August 2016]

Oxford University Press (2016h) *How Do You Decide If New Words Should Enter Oxford Dictionaries?* [online] available from <http://www.oxforddictionaries.com/words/how-do-new-words-enter-oxford-dictionaries> [22 August 2016]

Third Door Media (2014) *Search Engine Land – Google Blog Search Now Within Google News Search* [online] available from <http://searchengineland.com/google-blog-search-now-within-google-news-search-202202> [19 July 2016]

Urban Dictionary (2013) [online] available from <http://www.urbandictionary.com/> [16 November 2013]

Wiktionary (2016c) *Wiktionary: Criteria for Inclusion* [online] available from <http://en.wiktionary.org/wiki/Wiktionary:Criteria_for_inclusion> [22 August 2016]

Wiktionary (2016e) *Wiktionary: Entry Layout* [online] available from <https://en.wiktionary.org/wiki/Wiktionary:Entry_layout> [31 October 2016]

Wiktionary (2016f) 'How are Policies Enforced' *Wiktionary: Policies and Guidelines* [online] available from <https://en.wiktionary.org/wiki/Wiktionary:Policies_and_guidelines> [31 October 2016

Wiktionary (2016g) *Wiktionary: Welcome Newcomers* [online] available from <https://en.wiktionary.org/wiki/Wiktionary:Welcome,_newcomers> [31 October 2016]

Appendices

Appendix 1 – Job Advertisement: Senior Editor/Journalist

Senior Editor/Journalist

The candidates should be tech-savvy with a flair for writing interesting news stories in plain English for a time-poor audience. Candidates should demonstrate a passion for problems solving and an ability to adapt to changing scenarios in a proactive and pragmatic way. Candidates should also have an interest in feature articles, have experience in conducting interviews and an ability to understand what makes our news content important and usable for our audience. candidates should have some journalistic experience.

Duties:

- Writing news items, fixed to the needs of specific audiences on the latest legal/tax/accounting and regulatory developments on the day they are issued
- Edit, proof and rewrite copy to improve readability and ensure consistency
- Verify facts, dates and details using standard reference sources, including Lexis®Library
- Develop story or content ideas relevant to our readers or audience
- Monitoring and tracking editorial output to ensure coverage completeness

Qualifications:

- University graduate - preferably English/Politics/History/Journalism/Law/Accountancy or qualified by experience in the B2B/Professional publishing industry
- Broad understanding of legal and tax markets - key players.
- Good understanding of UK legal system
- Interest in current affairs
- Previous publishing experience
- Evidence of being able to deliver to tight and demanding timescales on multiple project
- Familiarity with Past common publishing/ XML/HTML publishing software
- Clear succinct writing style, ability to adapt to LN house style
- Influencing and collaboration - ability to influence and persuade others

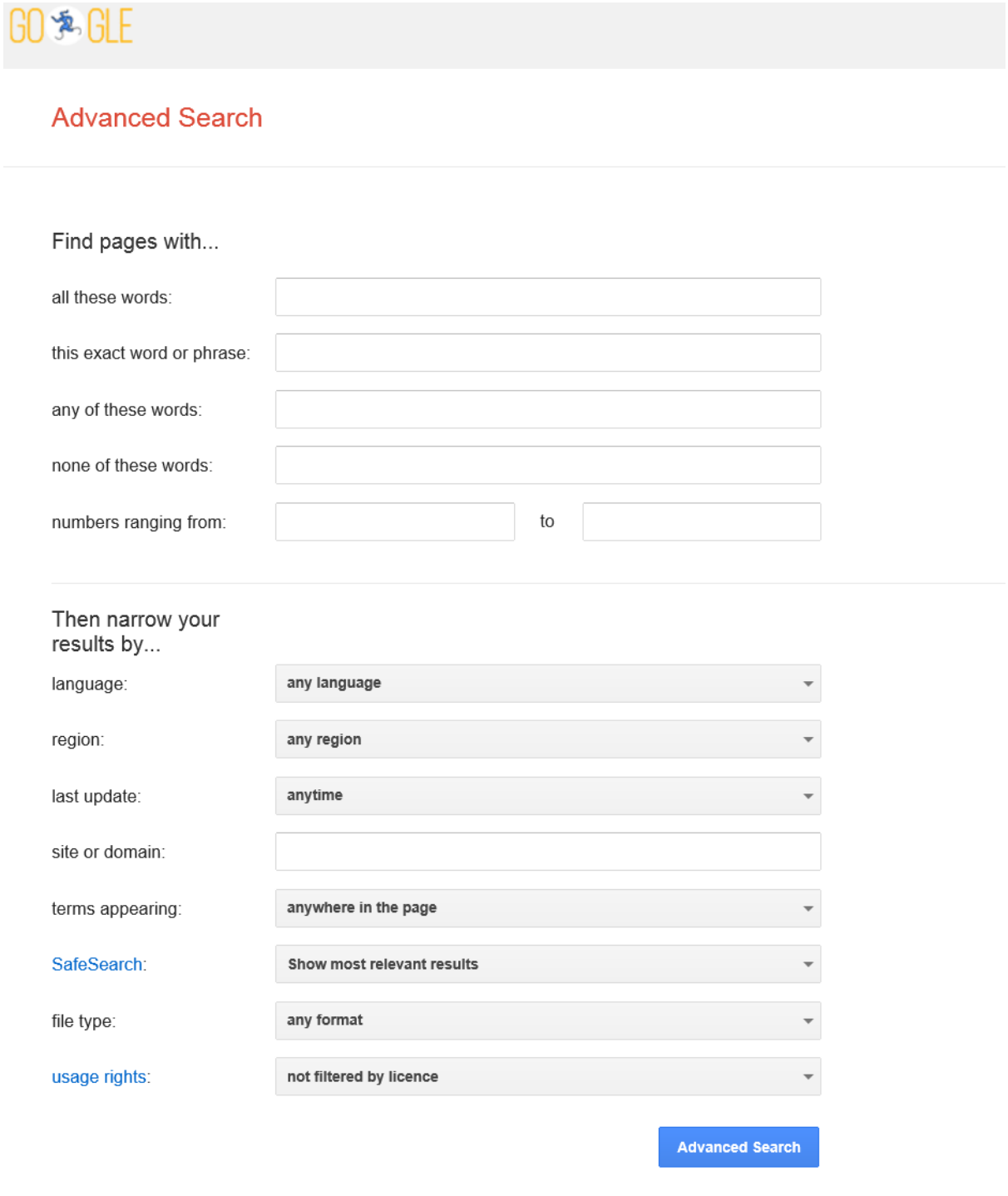
Excerpts from advertisement for a 'Senior Editor/Journalist' for LexisNexis, appearing in *The Guardian* online jobs section 15 July 2016: https://jobs.theguardian.com/job/6350457/senior-editor-journalist/

Appendix 2 – Blogs in *The Guardian*

Blog Name	Blog Archive
Comment is free	https://www.theguardian.com/uk/commentisfree
Technology blog	https://www.theguardian.com/technology/blog
Joe Public blog	https://www.theguardian.com/society/joepublic
Teacher's blog	https://www.theguardian.com/teacher-network/teacher-blog
Economics blog	https://www.theguardian.com/business/economics-blog+financial-crisis
Mortarboard blog	https://www.theguardian.com/education/mortarboard
Sportblog	https://www.theguardian.com/sport/blog
TV and radio blog	https://www.theguardian.com/culture/tvandradioblog+culture/radio
Music blog	https://www.theguardian.com/music/musicblog
Dave Hill blog	https://www.theguardian.com/uk/davehillblog
Apps blog	https://www.theguardian.com/technology/appsblog
Word of Mouth	https://www.theguardian.com/lifeandstyle/wordofmouth
Andrew Brown's blog	https://www.theguardian.com/commentisfree/andrewbrown
Higher Education Network blog	https://www.theguardian.com/higher-education-network/blog
Film blog	https://www.theguardian.com/film/filmblog
Books blog	https://www.theguardian.com/books/booksblog
Media blog	https://www.theguardian.com/media/media-blog
Datablog	https://www.theguardian.com/data
News blog	https://www.theguardian.com/news/blog
media monkey blog	https://www.theguardian.com/media/mediamonkeyblog
Hadley Freeman's blog	https://www.theguardian.com/commentisfree/hadley-freeman-blog
Social Enterprise blog	https://www.theguardian.com/sustainable-business/social-enterprise-blog
Environment blog	https://www.theguardian.com/environment/blog
Travel blog	https://www.theguardian.com/travel/blog
Developer blog	https://www.theguardian.com/info/developer-blog
Sustainability blog	https://www.theguardian.com/sustainability/blog
Organ Grinder	https://www.theguardian.com/media/organgrinder
Media Network blog	https://www.theguardian.com/media-network/media-network-blog
Poverty Matters blog	https://www.theguardian.com/global-development/poverty-matters
Work blog	https://www.theguardian.com/money/work-blog

Where 'blog' does not appear in the URL for the archive, it instead appears somewhere on the page of each entry. There are likely many more blogs than this, however since there is no directory of blogs, I can only include here the ones that I came across during this study

Appendix 3 – Commercial Search Engines Advanced Search Query Forms, As at September 2016



YAHOO! Yahoo - Search Home - Help

Advanced Web Search

You can use the options on this page to create a very specific search. Just fill in the fields you need for your current search.

Yahoo Search

Show results with

all of these words [ ] any part of the page
the exact phrase [ ] any part of the page
any of these words [ ] any part of the page
none of these words [ ] any part of the page

Tip: Use these options to look for an exact phrase or to exclude pages containing certain words. You can also limit your search to certain parts of pages.

Site/Domain

- (•) Any domain
- ( ) Only **.com** domains ( ) Only **.edu** domains
- ( ) Only **.gov** domains ( ) Only **.org** domains

( ) only search in this domain/site: [ ]

Tip: You can search for results in a specific website (e.g. yahoo.com) or top-level domains (e.g. .com, .org, .gov).

File Format

Only find results that are: all formats

SafeSearch Filter

Applies when I'm signed in:

- ( ) **Strict**: filter out adult web, video and image search results - SafeSearch On
- (•) **Moderate**: filter out adult video and image search results only - SafeSearch On
- ( ) **Off**: do not filter web results (results may include adult content) - SafeSearch Off

Note: Any user signed in on your computer as 18 or older can change this setting. We recommend periodically checking the SafeSearch Lock settings.

Advisory: Yahoo SafeSearch is designed to filter out explicit, adult-oriented content from Yahoo Search results. However, Yahoo cannot guarantee that all explicit content will be filtered out.

Learn more about protecting children online.

Tip: If you'd like to block explicit content for every search, you can set this in preferences. Keep in mind that this filter may not block all offensive content.

Country

any country

Languages

Search only for pages written in:

- (•) any language

OR

- ( ) one or more of the following languages (select as many as you want).

[ ] Arabic	[ ] French	[ ] Polish
[ ] Bulgarian	[ ] German	[ ] Portuguese
[ ] Chinese (Simplified)	[ ] Greek	[ ] Romanian
[ ] Chinese (Traditional)	[ ] Hebrew	[ ] Russian
[ ] Croatian	[ ] Hungarian	[ ] Slovak
[ ] Czech	[ ] Italian	[ ] Slovenian
[ ] Danish	[ ] Japanese	[ ] Spanish
[ ] Dutch	[ ] Korean	[ ] Swedish
[ ] English	[ ] Latvian	[ ] Thai
[ ] Estonian	[ ] Lithuanian	[ ] Turkish
[ ] Finnish	[ ] Norwegian	

Number of Results

Display 10 results per page.

Yahoo Search

YAHOO! News

News Home - Search Home - Yahoo! - Help

Welcome, Guest

Personalize News Home Page - Sign In

Your Search: Search

Save This News Search
View Your Saved News Searches
Advanced News Search

Web Directory News Yellow Pages Images NEW!

Use the drop down menus and radio buttons to change the parameters of your search.

Search Reset Search Tips | Help

Search Type:

News Stories | Matches on all words (AND)

Results Sorting and Display:

By Date and Relevance | Display 20 matches per page

Dates: (Search within a certain time period or specify your own)

Search only stories added during the past:

- (•) 30 days
- ( ) 1 day
- ( ) 3 days
- ( ) 1 week
- ( ) 2 weeks

( ) Specify a date or date range

Date format: use mm/dd/yy
For date range, use mm/dd/yy-mm/dd/yy

Sources: (A default search includes all sources; use this section to narrow by key sources)

News: see all sources

- [ ] AP
- [ ] Reuters
- [ ] USA TODAY
- [ ] Business Week
- [ ] E! Online
- [ ] The New York Times (free)
- [ ] Forbes
- [ ] AFP

Finance and Press Releases: see all sources

- [ ] Internet Wire
- [ ] Business Wire
- [ ] PrimeZone
- [ ] Canada NewsWire
- [ ] PR Newswire
- [ ] SmartMoney.com
- [ ] U.S. Newswire

Categories: (Choose as many as you want; default searches all categories)

- [ ] Top Stories
- [ ] World
- [ ] Politics
- [ ] Entertainment
- [ ] Business
- [ ] Crimes and trials
- [ ] Health
- [ ] Local
- [ ] Oddly Enough
- [ ] Community
- [ ] Commentary
- [ ] Science
- [ ] Sports
- [ ] Technology
- [ ] Elections
- [ ] Finance

Search

Web Directory News Yellow Pages Images NEW!

Appendix 4 – Low Risk Research Ethics Approval Checklist

Low Risk Research Ethics Approval Checklist

Applicant Details

Project Ref:	P16444
Full name:	Sharon Creese
Faculty:	[BES] Business, Environment and Society
Department:	[EL] English & Languages Dept
Module Code:	BESR010
Supervisor:	Hilary Nesi
Project title:	An Exploration into the Relationship between Lexicography and Language Growth in the Age of the Collaborative 'Wiki' Dictionary
Date(s):	21/01/2013
Created:	11/10/2013 13:04

Project Details

Examining real-world usage of new words as compared with Wiktionary definitions, in order to understand the impact of collaborative dictionaries on language growth occasioned by the legitimization and development of neologisms. Through this, determining whether Wiktionary now serves as an early predictor of neologism success or failure, and the implications for traditional publishers.

Participants in your research

Questions	Yes	No
1. Will the project involve human participants?		X

Risk to Participants

Questions	Yes	No
2. Will the project involve human patients/clients, health professionals, and/or patient (client) data and/or health professional data?		X
3. Will any invasive physical procedure, including collecting tissue or other samples, be used in the research?		X
4. Is there a risk of physical discomfort to those taking part?		X
5. Is there a risk of psychological or emotional distress to those taking part?		X
6. Is there a risk of challenging the deeply held beliefs of those taking part?		X
7. Is there a risk that previous, current or proposed criminal or illegal acts will be revealed by those taking part?		X
8. Will the project involve giving any form of professional, medical or legal advice, either directly or indirectly to those taking part?		X

Risk to Researcher

Questions	Yes	No
9. Will this project put you or others at risk of physical harm, injury or death?		X
10. Will project put you or others at risk of abduction, physical, mental or sexual abuse?		X
11. Will this project involve participating in acts that may cause psychological or emotional distress to you or to others?		X
12. Will this project involve observing acts which may cause psychological or emotional distress to you or to others?		X
13. Will this project involve reading about, listening to or viewing materials that may cause psychological or emotional distress to you or to others?		X
14. Will this project involve you disclosing personal data to the participants other than your name and the University as your contact and e-mail address?		X
15. Will this project involve you in unsupervised private discussion with people who are not already known to you?		X
16. Will this project potentially place you in the situation where you may receive unwelcome media attention?		X
17. Could the topic or results of this project be seen as illegal or attract the attention of the security services or other agencies?		X
18. Could the topic or results of this project be viewed as controversial by anyone?		X

Informed Consent of the Participant

Questions	Yes	No
19. Are any of the participants under the age of 18?		X
20. Are any of the participants unable mentally or physically to give consent?		X
21. Do you intend to observe the activities of individuals or groups without their knowledge and/or informed consent from each participant (or from his or her parent or guardian)?		X

Participant Confidentiality and Data Protection

Questions	Yes	No
22. Will the project involve collecting data and information from human participants who will be identifiable in the final report?		X
23. Will information not already in the public domain about specific individuals or institutions be identifiable through data published or otherwise made available?		X
24. Do you intend to record, photograph or film individuals or groups without their knowledge or informed consent?		X
25. Do you intend to use the confidential information, knowledge or trade secrets gathered for any purpose other than this research project?		X

Gatekeeper Risk

Questions	Yes	No
26. Will this project involve collecting data outside University buildings?		X
27. Do you intend to collect data in shopping centres or other public places?		X
28. Do you intend to gather data within nurseries, schools or colleges?		X
29. Do you intend to gather data within National Health Service premises?		X

Other Ethical Issues

Questions	Yes	No
30. Is there any other risk or issue not covered above that may pose a risk to you or any of the participants?		X
31. Will any activity associated with this project put you or the participants at an ethical, moral or legal risk?		X

Other Documents submitted
